NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AUGMENTING COMPREHENSION OF SPEECH IN NOISE WITH A COMPUTER-GENERATED FACIAL AVATAR AND ITS EFFECT ON PERFORMANCE

by

William R. Swann

December 2010

Thesis Advisor:	Lawrence G. Shattuck
Second Reader:	Michael E. McCauley

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** December 2010	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE Augmenting Comprehension of Speech in Noise with a Facial Avatar and Its Effect on Performance		**5. FUNDING NUMBERS**	
6. AUTHOR(S) William R. Swann			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A		**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number NPS2010.0112-IR-EP7-A.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited		**12b. DISTRIBUTION CODE** A	

13. ABSTRACT (maximum 200 words)

Military operations often occur in noisy environments, which can interfere with effective verbal communication. Previous studies have established the effectiveness of allowing a listener to see the speaker's mouth. This study examined the efficacy of incorporating a computer-animated facial avatar into a visual display in order to improve the comprehension of speech in noisy environments, while performing concurrent tasks. It also examined the effect of the avatar on the performance of concurrent auditory and visual tasks.

Twenty volunteers participated in an experiment measuring verbal comprehension, concurrent task performance and gaze dwell times while auditory, verbal and visual tasks were being performed under noisy conditions. The results indicated that the simple presence of the facial avatar did not significantly improve verbal comprehension while performing concurrent tasks. However, the facial avatar significantly improved verbal comprehension when the tasks being completed concurrently were more difficult and/or auditory-type tasks. The participants' performance for the concurrent tasks was not significantly affected by the presence of the facial avatar. The incorporation of computer–animated facial avatars into visual displays has the potential to improve verbal comprehension in noisy environments, depending on the nature of the concurrent task.

14. SUBJECT TERMS Speech, Noise, Communication, Auditory, Visual, Verbal, Bimodal, Multimodal, Task, Performance, Facial Avatar,			**15. NUMBER OF PAGES** 97
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

AUGMENTING COMPREHENSION OF SPEECH IN NOISE WITH A COMPUTER-GENERATED FACIAL AVATAR AND ITS EFFECT ON PERFORMANCE

William R. Swann
Captain, Canadian Forces
B.S., University of Lethbridge, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN HUMAN SYSTEMS INTEGRATION

from the

NAVAL POSTGRADUATE SCHOOL
December 2010

Author: William R. Swann

Approved by: Lawrence G. Shattuck
Thesis Advisor

Michael E. McCauley
Second Reader

Robert F. Dell
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Military operations often occur in noisy environments, which can interfere with effective verbal communication. Previous studies have established the effectiveness of allowing a listener to see the speaker's mouth. This study examined the efficacy of incorporating a computer-animated facial avatar into a visual display in order to improve the comprehension of speech in noisy environments, while performing concurrent tasks. It also examined the effect of the avatar on the performance of concurrent auditory and visual tasks.

Twenty volunteers participated in an experiment measuring verbal comprehension, concurrent task performance and gaze dwell times while auditory, verbal and visual tasks were being performed under noisy conditions. The results indicated that the simple presence of the facial avatar did not significantly improve verbal comprehension while performing concurrent tasks. However, the facial avatar significantly improved verbal comprehension when the tasks being completed concurrently were more difficult and/or auditory-type tasks. The participants' performance for the concurrent tasks was not significantly affected by the presence of the facial avatar. The incorporation of computer–animated facial avatars into visual displays has the potential to improve verbal comprehension in noisy environments, depending on the nature of the concurrent task.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ANOVA	Analysis of Variance
ANR	Active Noise Reduction
C2	Command and Control
C4ISR	Command, Control, Communication, Computers, Intelligence, Surveillance and Reconnaissance
cm	Centimeter
dB	Decibel
fMRI	Functional Magnetic Resonance Imaging
HINT	Hearing in Noise Test
HL	Hearing Level
HMD	Helmet-Mounted Display
HUD	Head-Up Display
Hz	Hertz
OSHA	Occupational Safety and Health Administration
mm	Millimeter
Phoneme	The smallest unique segment of speech used to form a spoken language
Pinnae	The external portion of the ears
PTT	Push-to-Talk
PUHLES	Physical capacity or stamina (P), Upper extremities (U), Lower extremities (L), Hearing and ears (H), Eyes (E), and Psychiatric (S) (Factors of the Military Physical Profile Serial System)
SA	Situational Awareness
S/N	Speech and Noise
SNR	Signal-to-Noise Ratio
SPIN	Speech Perception in Noise
SPL	Sound Pressure Level
SRT	Speech Reception Threshold
TTC	Talk-Through-Circuitry
Viseme	The visual portion of a phoneme

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

As early as the 1950s, researchers recognized that the ability to view the face of a speaker improved the comprehension of speech, especially in a noisy environment. There should be a means to capitalize on this phenomenon to improve the efficacy of communication in military environments. This improvement should be achievable through the presentation of a computer-animated facial avatar that provides the visual portion of phonemes to supplement the auditory component of verbal communication. Such an avatar could potentially be incorporated into a head-up display (HUD) or a helmet-mounted display (HMD).

Although it would be most desirable to provide the listener with a live video feed of the speaker, this presentation would require a great deal of bandwidth to do so. Using software to generate an animated face in the listener's display negates the need for a camera and does not require extra bandwidth; any radio signal could be used to generate the avatar. Advances in computer processing power and memory capacity tend to occur at a much higher rate than improvements in bandwidth and data compression.

The primary goal of the present research is to determine whether the presentation of a computer-animated facial avatar increases comprehensibility of speech-in-noise while participants are performing concurrent tasks. A secondary goal is to determine whether the presentation of a computer-animated facial avatar alters the performance of the concurrent tasks. Therefore, the hypothesis being investigated is as follows: the use of a computer-animated facial avatar will improve performance in a multitask scenario that requires multimodal processing (visual and auditory).

In order to determine the efficacy of the facial avatar, it was necessary to incorporate it into a series of visual and auditory tasks. There were four independent variables with two levels each: Speech Modality (facial avatar/no

facial avatar), Sentence Predictability (high/low), Task Type (auditory/visual) and Task Difficulty (high/low). This resulted in a 2 x 2 x 2 x 2 factorial design. The dependent variables being measured were Word Identification (comprehension of speech-in-noise), Task Performance (performance on concurrent tasks) and Gaze Dwell Time (time participants focused on the avatar).

Twenty volunteers participated in a series of tasks that each had a verbal subtask and either a visual or an auditory subtask. The results indicated that the simple presence of the facial avatar did not significantly improve verbal comprehension while performing concurrent tasks. However, the facial avatar significantly improved verbal comprehension when the tasks being completed concurrently were more difficult and/or auditory-type tasks. The participants' performance for the concurrent tasks was not significantly affected by the presence of the facial avatar. The incorporation of computer–animated facial avatars into visual displays has the potential to improve verbal comprehension in noisy environments, depending on the nature of the concurrent task.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my wife, Jacqueline, for her patience and support.

I would like to express my sincere gratitude to my thesis advisor, Dr Lawrence G. Shattuck, and my second reader, Michael E. McCauley. Without their guidance and advice this thesis would not have been possible. Their dedication to the students of the Naval Postgraduate School does not go unappreciated.

I am compelled to thank my Human Systems Integration classmates for their assistance in fine-tuning the experimental tasks. “Do you have a few minutes to try this out?” was always answered in the affirmative.

The Human Systems Integration research assistant, Diana Kim, deserves mention as well. Without her, the lab would not operate nearly as well.

Lastly, I wish to thank those Naval Postgraduate School students who took time out their busy schedules to volunteer to participate in the experiment.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. OVERVIEW

As early as the 1950s, research recognized that the ability to view the face of a speaker improves the comprehension of speech, especially in a noisy environment. Given the improved performance that results from combining the visual modality with the auditory modality, there ought to be a means to capitalize on the efficacy of bimodal communication channels in military environments. This improvement may be achievable through the presentation of an animated facial avatar that provides the visual portion of phonemes to supplement the auditory component of verbal communication. Such an avatar could potentially be incorporated into a head-up display (HUD) or a helmet-mounted display (HMD). It would be most desirable to provide the listener with a live video feed of the speaker; however, this would require a great deal of bandwidth and a camera would need to be aimed at the mouth of the individual speaking. Using software to generate an animated face in the listener's display negates the need for a camera and does not require extra bandwidth; any radio signal could be used to generate the avatar. Advances in computer processing power and memory capacity are occurring at a much higher rate than advances in bandwidth and data compression.

Previous studies (e.g., Sumby & Pollock, 1954; Summerfield, 1992; Massaro & Cohen, 1995) have established that the presentation of visemes (a term sometimes used to describe the visual component of phonemes) improves the perception of speech in noisy environments. It has also been established that computer-generated faces are also effective in improving perception of speech (Massaro & Cohen, 1995). There is surprisingly little research available regarding the combination of computer-generated faces and noisy environments. The current trends regarding the use of visemes to supplement audio phonemes are: to interpret the visemes to aid in filtering noise from the audio signal (Girin,

Schwartz, & Feng, 2001) and to incorporate visemes into speech recognition software (Nefian, Liang, Pi, Liu, & Murphy, 2002).

The literature review did not reveal any research attempting to determine if the presentation of the visual component of phonemes would either act as a distraction (increase workload/decrease performance) or be ignored by individuals undertaking a concurrent task.

B. BACKGROUND

"I see what you are saying." On face value, this statement is an amusing contradiction of two different senses used to indicate understanding of another person's point of view. However, upon further consideration, there is a deeper truth to this statement. Subconsciously, individuals rely on the visual components of speech as part of their daily lives (Massaro & Cohen, 1995). Anecdotally, there are individuals that claim to hear the television better with their glasses on and many individuals express a strong dislike for poorly dubbed foreign films. Children born blind tend to develop some aspects of speech more slowly than sighted children. In addition to "bleeping" or blanking the sound of censored words, network television producers routinely cover or blur the mouths of individuals using offensive language. These are but a few of the examples that illustrate the bimodal nature of verbal communication in daily life.

Communication in a military setting is often restricted to the auditory modality only; under ideal conditions the absence of the visual aspect of speech does not substantially affect the accurate communication of information. Unfortunately, the military does not operate only when conditions are ideal; noise is a common barrier to effective communication. It may even be argued that the noisiest situations are the ones in which effective communication is the most important. As early as the 1950s, experts have suggested that the inclusion of visual cues to augment auditory communication in noisy environments, including military environments, to improve the intelligibility of oral speech (Sumby & Pollack, 1954).

C. OBJECTIVE

This thesis is intended to act as a "proof of concept" regarding the implementation of computer-generated facial avatars into displays (such as HUDs and HMDs) in potentially noisy military environments. Improvements in the comprehension of verbal communication should have a positive impact on the execution of missions.

The primary goal of the present research is to determine whether the presentation of a computer-animated facial avatar increases comprehensibility of speech-in-noise while participants are performing concurrent tasks. A secondary goal is to determine whether the presentation of a computer-animated facial avatar alters the performance of the concurrent tasks. Therefore, the hypothesis being investigated is as follows: the use of a computer-animated facial avatar will improve performance in a multitask scenario that requires multimodal processing (visual and auditory).

D. RELEVANT DOMAINS OF HUMAN SYSTEMS INTEGRATION

Human Factors: The psychological processes involved in the integration of the auditory portion of speech with the visual cues can be classified as cognitive ergonomics. Workload, attention, and human performance are typically considered within the human factors domain (Licht, Polzella, & Boff, 1989).

Safety: Improved communication through increased comprehension of speech has the potential to improve safety. However, there is also the concern that an individual may become distracted by an animated facial avatar.

Personnel: If visual cues improve the comprehensibility of speech in noisy environments, it stands to reason that these visual cues may also act to compensate for hearing loss in individuals using equipment that could incorporate displays with facial avatars. This may act in a similar way as corrective lenses for individuals with visual deficiencies, increasing the number of personnel available for roles from which they might otherwise be excluded.

Training: Improved communication during training (especially in noisy environments) should increase the students' comprehension of their instructors' directions, increasing the effectiveness of the training session.

II. LITERATURE REVIEW

A. IMPORTANCE OF COMMUNICATION

C4ISR (Command, Control, Communication, Computers, Intelligence, Surveillance and Reconnaissance) involves the collection, use and dissemination of information. Although command and control (C2) are the most important activities associated with C4ISR, superior communication is required to enable commanders to exercise control of their resources (Department of Defense, 2010). Communication can, in simplest terms, be considered to be the process of transferring information. Information, in turn, has two basic uses; the first is to improve situational awareness (SA) in order to facilitate decision making and the second is to allow commanders to coordinate the implementation of those decisions. Situational awareness can be defined as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995).

B. INFORMATION QUALITY

Information quality can be assessed in terms of the following key criteria: accuracy, relevance, timeliness, usability, completeness, brevity and security (Department of Defense, 2010). This thesis focuses on the following five of those seven criteria. Accuracy refers to the degree to which the information received conveys the true situation and the receiver correctly interprets the message of the sender. Timeliness refers to the reception of the message in time to make decisions and act on the information; repetition of the message reduces timeliness. Usability refers to the message being understandable, in a commonly understood format. Completeness refers to the comprehensiveness of the information, the degree to which the entire message is articulated,

transmitted and received. Brevity typically refers to ensuring that the amount of information is kept to a minimum; it also refers to reducing unnecessary repetition of the message.

Dyer and Tucker (2009) found that the leaders they surveyed stressed the importance of verbal communication and the maintenance of situational awareness. They also determined that the voice communication function of the Land Warrior system was the most used component of the system, used by 84% of the leaders and 69% of the non-leaders.

Maximizing information quality avoids the need to add unnecessary complexity and contributes to the successful completion of the receivers' activities. The completion of other tasks is made more difficult when communication quality is degraded.

C. NOISE AS A BARRIER TO COMMUNICATION

At high intensity levels, noise can temporarily or permanently impair hearing or otherwise interfere with verbal communication; at lower intensity levels, noise may still interfere with the comprehension of verbal communication.

Phonemes are the smallest unique segment of speech used to form spoken language. These phonemes have both an auditory and visual aspect. Although phonemes are technically associated with both visual and auditory modalities, in practice visemes refer to the visual portion of speech while phonemes commonly refer to only the auditory portion. Visemes are comprised of the movement and shapes made by the face, predominantly the lips but the visualization of the teeth and tongue contribute to a lesser extent as well.

Each viseme may be associated with more than a single phoneme. For instance, the mouth makes a similar shape to produce both the "m" and "p" despite their dissimilar sounds (Lucey, Martin, & Sridharan, 2004). In other cases, two similar sounds may have very different visemes associated with them; such as "m" and "n." In the English language, the 48 most common phonemes

are associated with 14 visemes; these auditory and visual cues are used together in face-to-face conversations to facilitate communication. The multimodal nature of human speech allows individuals to communicate fairly effectively; even when one modality is impaired the other can compensate. However, when both modalities are impaired the effectiveness of communication suffers. Communicating with speech remotely (e.g., by radio) when the listener is in a noisy environment is an example of both modalities being impaired; thus communication effectiveness is often reduced.

D. COMMON METHODS FOR IMPROVING COMMUNICATION IN NOISY ENVIRONMENTS

Noise in the sender's environment can be counteracted by using a noise-cancelling microphone. This technique is basically a subtraction method. The microphone the sender speaks into picks up both the words spoken and the noise in the environment, a second microphone picks up the environmental noise; the signal transmitted to the receiver is the message plus the noise, minus the noise.

A more innovative solution that was investigated involved using a video camera to detect the speaker's mouth movements to predict the phoneme spoken (Girin, Schwartz, & Feng, 2001). A complex algorithm was then employed to enhance the spoken message while filtering out the environmental noise. Although this method was designed to improve the listener's comprehension of speech when the sender was in a noisy environment, it demonstrated that the visual and auditory aspects of speech can be combined to enhance a message's comprehensibility.

Noise in the recipient's environment can be counteracted by simply increasing the volume of speakers/headphones, employing passive noise reducing headphones or employing active noise reducing headphones. The approach of merely increasing the decibel level of the radio in comparison to decibel level of the environmental noise may have utility when the noise is only

moderately loud, but has inherent limitations. Speakers/headphones have limited maximum volumes (decibel levels); therefore it may not always be possible to increase the volume high enough to make the message understandable. In some cases, the noise and radio volumes combined may be high enough to cause either temporary or permanent hearing damage. Also, as the volume increases the message may become distorted because of the (poor) quality of the speakers or the headsets. As well, messages are often clipped to save on bandwidth.

Employing passive noise reduction—either through the use of headphones (earmuffs), earplugs or both (dual protection)—acts to reduce the intensity of the environmental noise reaching the listener's ears. But this noise reduction technique may also diminish important sounds in the environment such as alarms, important changes to engine noise or other cues from the environment that would otherwise serve to increase awareness (Abel, Tsang & Boyne, 2007).

Earplugs can be designed to attenuate noise in either a linear or nonlinear manner. Linear earplugs attenuate noise relatively constantly across all audible frequencies. Nonlinear earplugs attenuate more noise at lower frequencies. Passive noise reduction headphones are also more efficient at attenuating low frequency noise than high frequency noise. The advantage of nonlinear noise attenuation is that it reduces the intensity of low frequency noises (such as those generated by vehicles, aircraft, heavy machinery and weapons) more than it reduces higher frequency sounds (such as those generated by speech and alarms).

E. PROBLEMS ASSOCIATED WITH ACTIVE NOISE REDUCTION

Active noise reduction (ANR) acts to reduce the intensity of the environmental noise reaching the listener's ears by using a technique similar to that of a noise-cancelling microphone. The headset worn by the listener utilizes a microphone to sample the environmental sounds, and then transmits it to the

listener 180 degrees out of phase as destructive interference, thereby cancelling out (or at least greatly reducing) the environmental noise and presenting only the intended message. Active noise reducing headsets are most efficient at reducing repetitive (i.e., periodic) noise due to its more predictable nature. And, like passive noise reduction headsets, active noise reduction headsets tend to attenuate low frequency sounds more than, high frequency sounds.

However, short-term abrupt onset noise (i.e., impulse noise, such as weapon noise) can bypass the protection normally associated with both passive and active noise reducing headsets. Impulse noise can cause a "ringing" within the headset due to repeated compression and rarefaction cycles within the protective cups. Impulse noises can be particularly troublesome with headsets employing ANR due to its reactive nature; the attempted cancellation of a single impulse can induce "ringing" that may reach decibel levels higher than measured in passive noise reduction headsets (Buck, 2000). It should be noted that earplugs dampen impulse noises without ringing.

More advanced headsets have "talk-through" capabilities that electronically amplify the ambient sounds that are below a pre-established threshold, but still attenuate high intensity sounds.

F. IMPORTANCE OF ENVIRONMENTAL SOUNDS

The indiscriminate use of hearing protection, either active or passive, may result in a diminished level of situational awareness. In an aircraft, changes in engine sounds or wind noise provide valuable information as to the aircraft's condition. To an individual in a potentially hostile outdoor setting, the twigs snapping, changes to bird sounds, vehicles/aircraft approaching or other unusual sounds provide useful information about the surrounding area (Scharine, Henry, & Binseel, 2005). As important as hearing protection is, when sound levels reach the threshold of either temporary or permanent hearing damage, the

inappropriate use of hearing protection in mildly noisy environments, simply to improve radio communication, may reduce situational awareness to an unacceptable level.

Beyond simply attenuating ambient sounds, hearing protection tends to disrupt the ability of individuals to determine the location of a sound (Abel et al., 2007). Sound localization is achieved through a combination of several means: the difference in loudness of the sound as it reaches each ear, the difference in the amount of time it takes to reach each ear, and the shape of the pinnae. The shape of the pinnae, the external portion of the ears, scatters incoming sounds in a manner that is unique to the direction from which the sound comes. Headphones have a more disruptive effect on sound localization than earplugs, due to the ear cups distorting the sound before it reaches the pinnae.

Abel and Paik (2005) suggest that headphones should not be worn when sound localization is an important component of the tasks to be performed. In a later study, Abel et al. (2007) determined that the use of active noise reducing headphones results in the poorest ability to localize sounds. Active noise reducing headphones produce left/right reversal of the sound localization more often than either passive noise reducing headphones or earplugs. The use of newer technologies, such as talk-though-circuitry (TTC) and push-to-talk (PTT), result in better sound localization than other sound attenuation devices; but still lead to errors in sound localization.

Hearing protection devices are necessary when noise levels approach a high enough intensity to causing damage. However, they have too many drawbacks to be used routinely as a method for improving the comprehensibility of speech in noisy environments when the potential for hearing damage is not present (i.e., when the sound pressure level is not likely to exceed safe levels).

G. VISUAL CUES AS AIDS TO COMPREHENSION OF SPEECH IN NOISY ENVIRONMENTS

If simply attenuating environmental sounds is not a universal solution to improving communication, another means to improve verbal communication needs to be explored. With the steady improvements in computer hardware and software technology, the traditionally audio form of communication may be supplemented with visual cues. Sumby and Pollack (1954) were among the first researchers to investigate the influence of the visual factors of speech on the listeners' comprehension of spoken words in a noisy environment. They compared their participants' ability to correctly identify words spoken in a noisy environment. The words were presented as the participants either faced toward or away from the speaker. Background noise (white noise) was held at a constant decibel level and the loudness of the speech was varied. Headphones were used to control the decibel levels of the noise and spoken words.

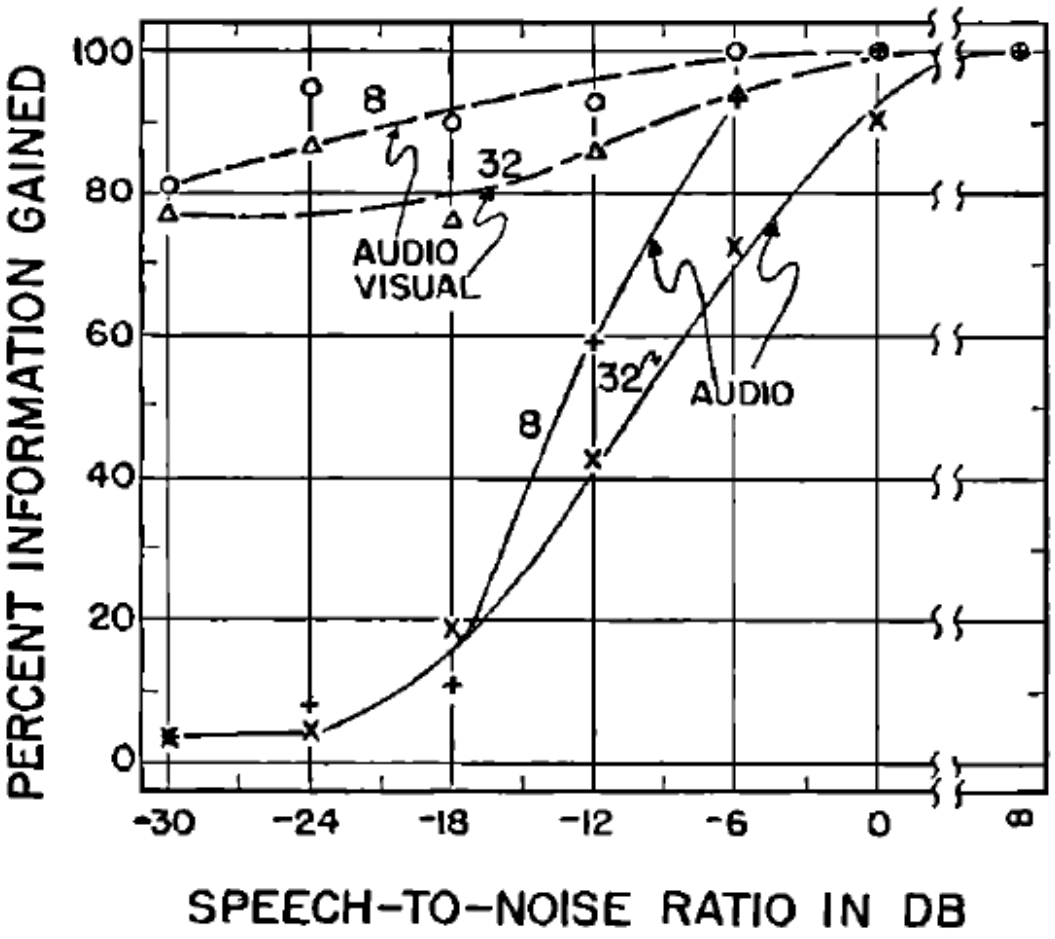

Figure 1. Speech intelligibility at various noise levels by audio and audio/visual pPresentation (From Sumby & Pollack, 1954)

The difference between the loudness of the speech and the noise (S/N) was varied from 0 dB (speech and noise at equal intensities) to –30 dB (the speech 30 dB quieter than the noise). Speech intelligibility was determined by tallying the number of correctly identified words. At all S/N levels tested, speech intelligibility scores were higher when the speakers face was visible to the listener.

As the loudness of the speech decreased relative to the noise, comprehension of speech decreased regardless of whether the words were presented with or without the visual component of speech (Figure 1). However,

as the speech-to-noise ratio decreased (i.e., became more negative) the audio/visual presentation performed increasingly better than the auditory only presentation of the words.

Summerfield (1992) examined the importance of lipreading. Lipreading tends to improve speech comprehension in noise to a degree equivalent to a 4–6 dB reduction in noise level. This equates to a 10–15% improvement in intelligibility of speech due to the incorporation of the visual aspect of speech. The researcher also determined that the additive effect of the visual component of speech is quite robust against desynchronization. The audio component may precede the visual component by up to 140 ms, or follow the visual component by up to 80 ms, before the benefit of integrating the auditory and visual component dissipates. This result indicates that the synchronization of a video image of words being spoken with the sound of speech does not need to be perfect.

Ross, Saint-Amour, Leavitt, Javitt, and Foxe (2007) further investigated the utility of allowing participants to view a face speaking in a noisy environment. They sought to better describe the inverse relationship between the helpfulness of bimodal communication and the intensity of the interfering noise. Early studies, such as the one by Sumby and Pollack (1954), provided the participants with the lists of words that would be spoken in the noisy environment, artificially facilitating the identification of the words spoken in the noisy surroundings. Not only did Ross et al. not pre-expose their participants to the words to be spoken, they also limited the words to be identified to monosyllabic words. This was done to ensure that partially comprehended words were not correctly identified based on clues within the word; i.e., "~~xxx~~cake" may be correctly guessed to be "cupcake."

Bimodal communication (auditory and visual) significantly improved comprehension at all noise levels, with a -12 dB SNR (signal-to-noise ratio) possessing the highest difference in comprehension rates between bimodal and auditory-only presentations of verbal communication (Figure 2; Ross et al.,

2007). An interesting demonstration of the synergistic effect of the auditory and visual aspects of verbal communication is that the percent of correct responses for the auditory-visual presentation of words exceeds correct responses from auditory-only and visual-only combined (when noise exceeds speech by six or more decibels).

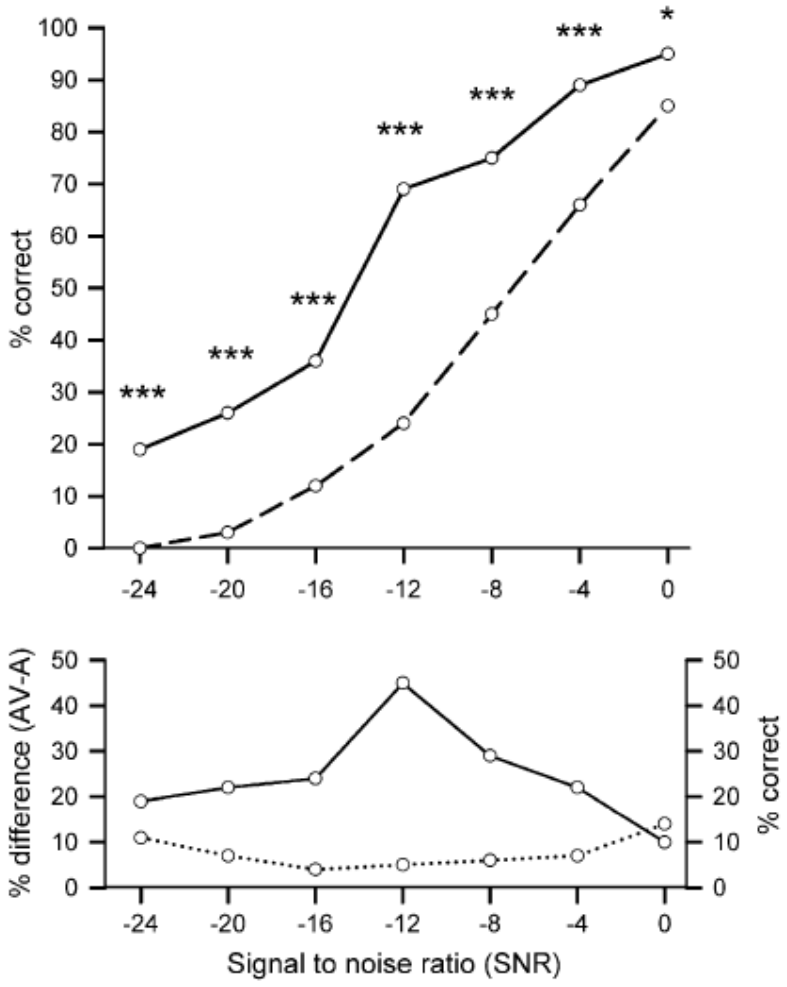

Figure 2. The top panel depicts the percentage of correctly identified words (% correct) depending on the SNR for the auditory-alone (A: dashed line) and the AV (solid line) conditions. Significant differences between both conditions are indexed with stars (*$p < 0.05$; ***$p < 0.001$). The bottom panel shows the multisensory gain as the difference (AV-A) in speech recognition accuracy as a function of level of SNR (solid line). The dotted line represents performance in pure speech-reading (V) in percent correct (From Ross et al., 2007)

Calvert et al. (1997) explored the neurological reaction to the visual component of speech. The silent viewing of a face mouthing words was found to activate auditory cortical sites in the brain that are normally activated while listening to spoken words. Simple mouth movements that did not resemble human speech (i.e., not resembling words being spoken) had virtually no effect; while the mouthing of words displayed similar stimulation observed during fMRI (functional Magnetic Resonance Imaging) examinations to a person actually listening to words being spoken. This study highlights the physiological basis for the augmentation of the auditory portion speech with visemes.

Calvert and Campbell (2003) continued their exploration of the neurological relationship between auditory and visual components of speech. In this study, participants were instructed to examine both still and moving images of a face with no audio and to indicate when the target phonemes were presented. When the correct visible phonemes were presented, fMRIs revealed that regions of the brain normally associated with auditory processing were activated. The presentation of images that did not match the target phonemes did not result in activation of those areas. The moving images resulted in higher levels of activation. This study revealed that the additive effect of the visual component of speech is not limited to observing moving images of a face mouthing words; still images of a face mouthing words also invoke a neurological reaction. These findings emphasize the impact of the visual aspect of speech on the processing of verbal communication.

The visual aspects of speech are not limited to the subjective processing of the human experience. Girin et al. (2001) explored using computer algorithms to integrate video images of lips with speech in noise to enhance comprehensibility of the speech. This technology used the movement of the speaker's lips to improve the quality of the audio signal sent to the listener. Although this technology removes noise that is present in the sender's environment rather than the listener's, it demonstrated that augmenting the audio

component of speech with the visual portion is not limited to the psychological realm, but the two modalities also can be objectively integrated.

Noise is not the only barrier to verbal communication that can be remedied by augmenting the auditory component with the visual component. Chen and Hazan (2009) investigated the efficacy of bimodal communication when interacting with non-native speakers. The “non-native speaker effect” suggests that native speakers of a language benefit (i.e., correctly interpret phonemes) when the visual components of speech are available while listening to non-native speakers.

The ability to see the speakers’ mouth movements benefits English speakers who may be required to listen to the non-native speakers (or to individuals with strong accents), thus improving their comprehension (Figure 3). This may prove useful in a military setting when the listeners are required to communicate with individuals whose first language is not the same, whether it is fellow countrymen or members of a multinational force.

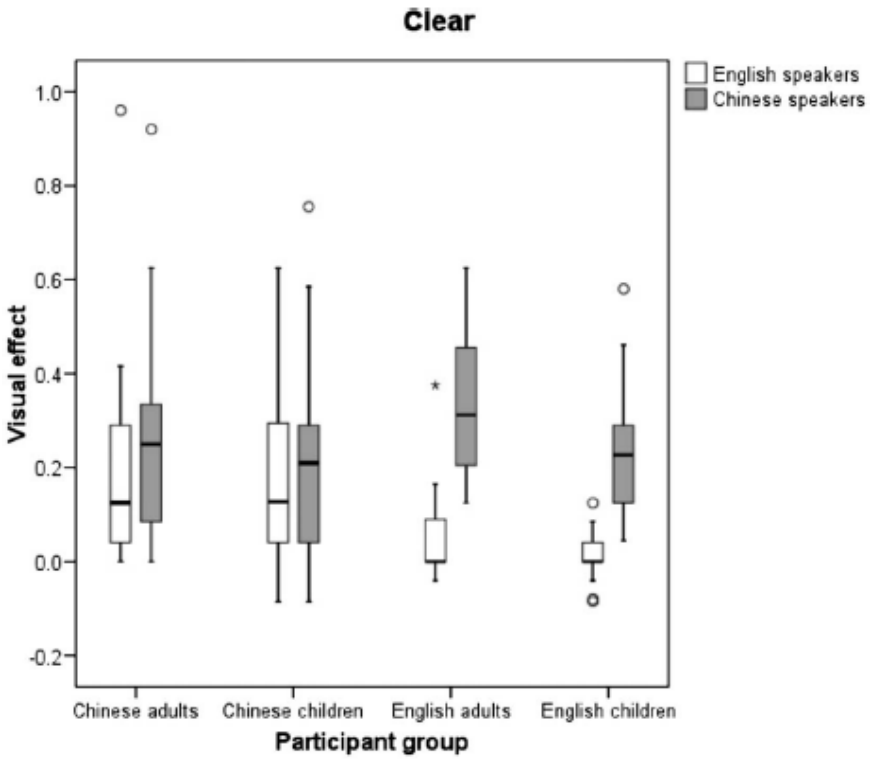

Figure 3. Native English-speaking adults were significantly influenced by the availability of the visual component of speech when listening to non-native English speakers (From Chen & Hazan, 2009)

Most studies focus on the comprehension of the English language using both auditory and visual cues. English is a Germanic language and is often referred to as a language of consonants. In contrast, French is a Romance language and is referred to as a language of vowels. Robert-Ribes, Schwartz, Lallouche, and Escudier (1998) investigated the effectiveness of presenting visual cues when French vowels were vocalized in a noisy environment. The researchers indicated that the augmentation of auditory cues with visual cues are most effective at a SNR of -12 dB; under all noise levels listeners correctly identified the phonemes presented to them more often when both auditory and visual aspects of the French vowels were provided. This indicates that the synergistic effect of bimodal communication is not restricted to the English language; the effect may be utilized for both English and French speakers. This

information may be especially useful in justifying the implementation of bimodal communication in bilingual armed forces, such as the Canadian Forces.

H. IMPRACTICALITY OF TRANSMITTING FULL VIDEO

Based on these studies, it would be ideal to transmit live audio/video moving images as a means of remote communication. As desirable as this prospect is, the hardware and bandwidth required is currently impractical. The sender would need to have a camera aimed at his or her face and supplemental lighting may be required. This requirement would prohibit its use in many situations. Although the Land Warrior is now an outdated system, it serves as a relevant example of bandwidth restrictions. The commonly used frequency bands restricted data transmission to 9600 bits per second (Zieniewicz, Johnson, Wong, & Flatt, 2002); a single image can take up to 75 seconds to be transferred and displayed. Even if bandwidth and transmission limitations are resolved, the transmission of live video may not be the best use of the limited resource.

While data transmission rates increase relatively slowly, memory size and computer processing speed are increasing at a faster pace. Software exists that can generate an animated face, the mouth of which moves to match the phonemes of the audio signal. No special hardware is required at the sender's location; the visual cues can be generated from any audio signal. A computer-animated "talking" face, or just a mouth, that presents visemes in conjunction with auditory phonemes, and is generated at the recipient's end, may effectively improve speech perception and comprehension when the recipient is in noisy surroundings.

I. EFFECTIVENESS OF COMPUTER-ANIMATED FACES

Massaro and Cohen (1995) utilized animated faces to represent the visual portion of the phonemes that were synchronized to the audio. When the auditory and visual components of the syllables were in agreement, the percentage of correct interpretations of the phonemes was highest (compared to both unimodal

speech or conflicting cues). They effectively demonstrated that using visemes to enhance comprehension of phonemes is not limited to the presentation of natural faces; computer-generated facial avatars presenting visemes also are effective.

The effectiveness of using a computer-generated facial avatar was compared to that of a recorded image of a "live" moving face (Ouni, Cohen, Ishak & Massaro, 2007). The effectiveness of presenting only the lips was also compared to presenting a full facial image. Participants were instructed to identify the phonemes presented to them under various levels of noise, as well as silent viewing of visemes. The visemes were presented as either a natural face, natural lips, synthetic face (i.e., computer-generated), synthetic lips or auditory only.

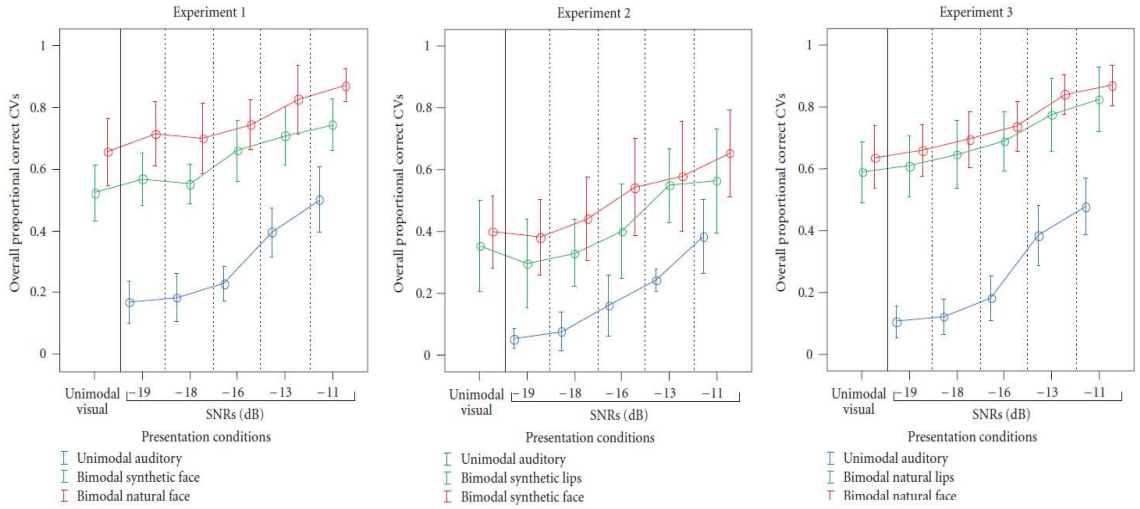

Figure 4. Proportion of correctly identified CVs (consonant-vowel phonemes) under various conditions; bars indicate one standard deviation (From Ouni et al., 2007)

Each of the three experiments revealed that the correct identification of the phonemes was significantly improved when supplemented with visual cues (Figure 4). Although the natural face significantly outperformed the computer-generated avatar, the computer-generated avatar still performed significantly better than the auditory-only presentation of the phonemes. The presentation of

the entire face did not significantly outperform just the lips. This result should prove useful; the presentation of a less complex avatar will minimize the computer processing power required to display the visemes intended to aid comprehension of speech in noise.

Nicholls, Searle, and Bradshaw (2004), with the knowledge that the right side of a speaker's mouth is more expressive, questioned whether or not being able to observe the right side of a speaker's mouth is important in the perception of speech. They found that viewing the right side of the mouth is more important to lip-reading than the left side, but lip-reading is most effective when the entire mouth can be viewed. Individuals attend more to the right side, or what they perceive to be the right side, of a speaker's mouth than the left side; this information may be useful in determining the optimum placement of a facial avatar. Positioning the avatar on the right side of the listener's field of view will place the right side of the avatar's mouth closer to the center of the viewer's field of view, potentially maximizing effectiveness of the avatar while minimizing distraction from other tasks.

J. WORKLOAD AND CROSS-MODAL INTERACTIONS

To date, the studies relating to the visual component of speech in noisy environments have concentrated exclusively on the perception of speech; the implications of workload have not been examined. There is very little question as to the efficacy of supplementing the auditory aspect of verbal communication with visual cues, but the practicality of using that augmentation needs to be addressed. Improvements in comprehension may come at the expense of the performance of concurrent tasks, thereby reducing the usefulness of improving communication. Conversely, concurrent tasks may distract the recipient from attending to the visemes presented.

The multiple resource theory seeks to explain the interaction and conflict between spatial and verbal processes (Wickens, 2002). It addresses cross-modal interactions as well as intra-modal interactions; tasks with auditory and

visual modalities should not interfere with each other as much as multiple tasks using the same modality (i.e., two tasks that both use the visual modality or two tasks that both use the auditory modality). Two visual tasks might not interfere with each other if one involves focal vision while the other involves ambient vision. It is also suggested that concentrating on a difficult or important task may interfere with other tasks regardless of the process and modality differences.

When visemes are presented, communication is changed from auditory/verbal only to both auditory/verbal and visual/verbal. Bimodal communication theoretically distributes the workload between the auditory and visual channels. This distribution should reduce the workload associated with those particular channels with respect to the communication task, but may increase the potential number of workload conflicts. However, since the visual aspects of speech use visual-verbal rather than visual-spatial resources, it may not utilize enough of the spatial perceptive and cognitive resources to substantially affect workload.

The multiple resource theory suggests that the addition of a second input modality will increase workload, but it will only interfere with performance if attention is overtasked and workload approaches overload. The extent to which workload, performance and comprehension of speech interact with the addition of visual cues needs to be investigated.

K. TESTS OF COMPREHENSION OF SPEECH IN NOISY ENVIRONMENTS

The intelligibility of speech in noisy environments can be measured by various means. Two methods for determining how well participants correctly identify verbal messages are the SPIN (Speech Perception in Noise) and the HINT (Hearing in Noise Test). Both approaches have merit, but neither is universally applicable.

The HINT is used to determine the Speech Reception Threshold (SRT) and was developed in Britain in the 1990s (Giguere, Laroche, & Vaillancourt,

2008). In an environment with 65 dB of noise, a verbal sentence is presented at increasingly louder sound levels until the subject can correctly repeat every word of the sentence. After twenty sentences have been completed an individual's threshold for correct speech comprehension in a noisy environment is determined. This technique relies on absolute, rather than relative, sound levels; an individual's hearing acuity affects the results. Each participant's hearing acuity must be determined and accounted for in order for the results to be pooled. Score is assessed in terms of the Speech Recognition Threshold, in dB.

The SPIN test involves having the participants listen to sentences that have been combined with noise, at predetermined Signal-to-Noise Ratios (SNRs) (Kalikow, Stevens, & Elliot, 1977). The participants are instructed to identify the last word of the sentence, which is always a monosyllabic word. There are two types of sentences, high and low predictability. High predictability sentences are composed in such a manner that the wording of the sentence provides clues to the last word. For example, in "The boat sailed across the bay" the words "boat," "sailed" and "across" tend to suggest such words as "lake," "sea," "pond" and "bay." Low predictability sentences are composed such that the wording of the sentence does not provide any clues to the last word. For example, in "John is talking about the bay," the words preceding the target word do not suggest the final word. SPIN tests are typically comprised of fifty sentences, with an equal amount of high and low predictability sentences. The hearing acuity of each participant is of relatively low importance since the noise and speech are controlled relative to each other rather than as absolute values. Score is assessed as the number of correct responses at any given SNR.

Of these two tests for the comprehension of speech in noise, the SPIN test is more suited for use in this study. In order to use the HINT, all participants must have their hearing tested to the nearest dB across several frequencies. Because the noise is presented at a set dB level and the sound level for the sentences is added to it at increasing levels, the sound could potentially become dangerously loud before the sentence is comprehended. The SPIN test does not

require any special hearing tests, the participants need only self-report that they possess normal hearing. Because SNR is the primary factor, any inaccuracies in self-reporting should be relatively inconsequential. The SPIN test also lends itself to simpler and more consistent scoring when variables are manipulated.

L. SUMMARY

High quality communication is an important factor when conducting military operations. Noise is a common barrier to communication, but traditional methods for combating noise have drawbacks. Previous studies have established that allowing listeners to observe a speaker's mouth improves the comprehension of speech in noisy surroundings. Computer-animated mouths have been revealed to be as effective as video images of an entire face at improving the comprehension of speech in noise.

No studies were found that investigated the effects of a computer-animated facial avatar on both speech comprehension and performance on concurrent tasks. The hypothesis of this study is as follows: the use of a computer-animated facial avatar will improve performance in a multitask scenario that involves multimodal processing (visual and auditory).

THIS PAGE INTENTIONALLY LEFT BLANK

III. METHOD AND EXPERIMENTAL DESIGN

A. OVERVIEW

In order to determine the efficacy of the facial avatar, it was necessary to incorporate it into a series of visual and auditory tasks at two difficulty levels. This was accomplished by developing a series of computer-based visual target detection and tone-change detection tasks. Regardless of the type of task presented, a verbal message was concurrently presented as either an auditory-only message or with the lip-synched facial avatar. Eye tracking equipment was employed to evaluate the duration of time the participants' gaze dwelled on the facial avatar.

B. PARTICIPANTS

Volunteers were solicited via the Naval Postgraduate School email system. Prior to this solicitation; approval was sought, and granted, by the Naval Postgraduate School's Institutional Review Board to ensure that the participants' rights were protected.

Participation was open to all students originating from countries with English as an official language. Additionally, the email indicted that participants were required to possess "normal" visual and auditory acuity.

For the purpose of this study, "normal" was defined as meeting the Military Physical Profile Serial System (PUHLES) standards of "H" Position (hearing) of 2 or better (audiometer average level for each ear at 500, 1000, 2000 Hz, or not more than 30 dB, with no individual level greater than 35 dB at these frequencies, and level not more than 55 dB at 4000 Hz; or audiometer level 30 dB at 500 Hz, 25 dB at 1000 and 2000 Hz, and 35 dB at 4000 Hz in better ear). "Normal" for the "E" Position (vision) was defined as of 2 or better (distant visual acuity correctable to not worse than 20/40 and 20/70, or 20/30 and 20/100, or

20/20 and 20/400). Potential participants were screened via an “eye chart” and an audiogram prior to participation to ensure that they meet these visual and auditory acuity requirements.

Twenty students volunteered and were considered suitable for participation, six females and fourteen males. The age of the participants was ranged from 19 to 24 years old.

To ensure the participants’ safety, sound pressure levels were limited. Participants were not exposed to noise for a cumulative time longer than ten minutes at A-weighted sound levels louder than 74 decibels. This exposure is far less than the maximum sound levels prescribed for an eight-hour exposure (85 dB equivalent A-weighted sound level). The maximum noise exposure limits adhered to were in accordance with OSHA Standard 1910.95, as mandated by the Department of Labor regulations.

Of the twenty volunteers, four were excluded from participating in the eye tracking portion of the study due to the wearing of glasses. Eyeglasses tend to occlude the eye tracker’s view of the participant’s eyes. One participant’s eye tracking data was rejected; he changed his body position while performing the tasks and his gaze became untrackable.

C. APPARATUS

1. Software

Several software programs were used to create the auditory, verbal and visual tasks. The sentences from the SPIN test were used to produce the verbal messages through the use of text-to-speech software. The lip-synched facial avatar was then produced using animation software for the verbal tasks. The visual tasks consisted of identifying target icons on a disruptive background. Background noise and tones for the auditory task were then generated for the auditory tasks.

a. Speech Generation

Verbal messages were taken from the SPIN test. The target word to be identified was the last word of each sentence. There were two types of sentences: those with high-predictability target words and those with low-predictability target words. High predictable sentences were designed to allow the listener to anticipate the target word, while low predictability sentences did not aid in the correct determination of the target word. To control the loudness, tempo and emphasis of words in the sentences, text-to-speech software was employed. Audio files were created using *AT&T Labs Natural Voices ® Text-to-Speech Demo* (http://www2.research.att.com/~ttsweb/tts/demo.php), and was accessed directly online at the AT&T website.

b. Facial Animation

The audio files were imported into animation software to produce the lip-synched facial avatar. *CrazyTalk v 6.0 Pro* (version 6.0.0611.1) automatically analyzed the audio files and detected the phonemes, then synchronized the movements of the model's mouth to produce matching visemes. A grayscale face was selected from amongst the included facial models, and each of the SPIN sentence audio files was imported and processed automatically. The movements of the facial avatar's mouth were then adjusted to ensure the correct visemes were selected and the timing of the movements matched the speech. Movie files were exported at a resolution of 600 by 800 pixels at 30 frames per second and were four seconds in length. It should be noted that the speech, and synchronized facial movements, began one second after each movie file started. This delay was designed to ensure that the verbal message occurred midway through each task.

Figure 5. CrazyTalk 6 interface displaying selectable visemes

Every SPIN sentence was produced as both a facial avatar movie file and as a "blank" grey movie file. This ensured that both the auditory-visual (facial avatar) and auditory-only (grey) presentation of the verbal messages would be equivalent in terms of loudness and quality.

Because the resultant movie files displayed the entire face of the avatar, the moving images needed to be cropped. *VidCrop Pro* (version 1.1.0.23) was employed to isolate the avatar's mouth. To produce movie files that did not display the facial avatar, the movies were cropped to display only a portion of the grey non-moving background. The movie files were generated at a resolution of 320 by 160 pixels as a frame rate of 29 frames per second, with the audio portion sampled at 22050 Hz.

c. Tone and Noise Generation

The white noise for the background noise and tones for the auditory tasks were generated using *Audacity* (version 1.2.6). The background noise was

generated as white noise. The tones for the auditory tasks were 1200 milliseconds in duration; after the first 600 milliseconds, the tone would shift either to a higher or a lower frequency.

The audio files were outputted at a sampling rate of 44100 Hz. Each file was five seconds in duration to match the length of the tasks presented during the experimental testing session.

d. Visual Search Targets

The visual task consisted of searching for target icons amongst a mix of target and non-target icons on a disruptive background. The target icons were silhouettes of military vehicles randomly placed around a computer screen. The vehicle silhouettes were selected from a series of characters available in a military font set freely downloaded from http://www.dafont.com/military-rpg.font.

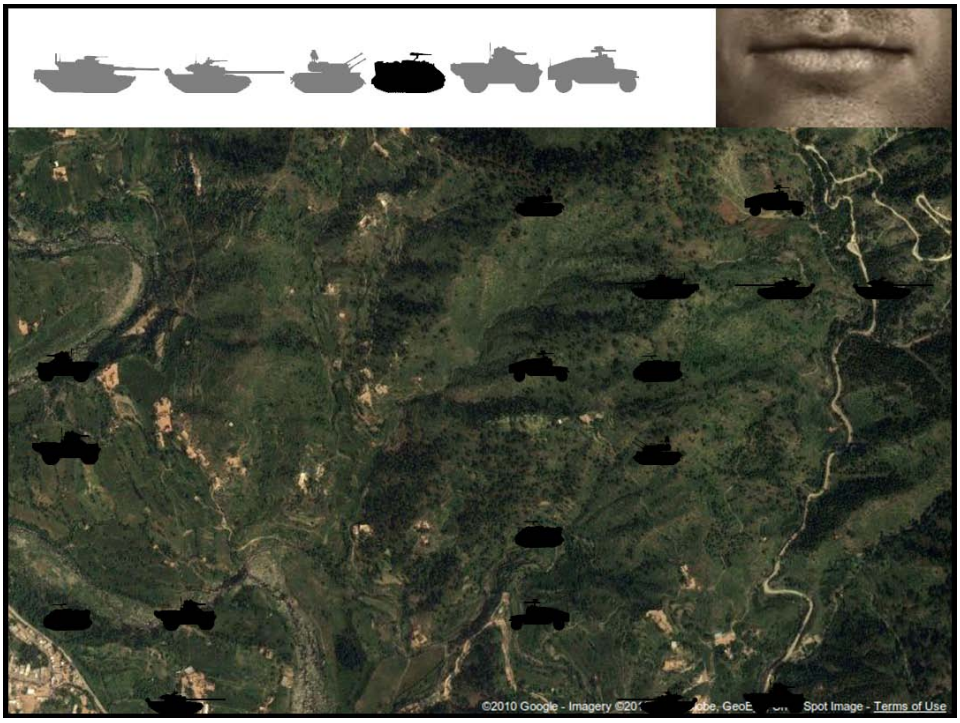

Figure 6. Example screenshot of a visual task

The background image was an aerial view of a forested mountainous area, obtained from *Google Earth*. The rationale for selecting a disruptive background image was that its disruptive nature increased the difficulty of differentiating between the various vehicle shapes. The combination of military vehicle silhouettes and the terrain background was intended to provide the added benefit of promoting a sense of working at a tactical console (Figure 6).

e. Data Analysis

The data collected during experimental testing was compiled using *Microsoft Excel 2007*, then imported in to *Minitab* (version 15.0) for analysis. Any differences in task performance scores were identified using ANOVA. Main and interaction effects with $p<0.05$ were considered significant.

2. Hardware

Although the experimental testing was performed on a single computer, several pieces of ancillary equipment were required for preparation and support. An eye tracking system recorded participant's gaze and headphones were used to reduce the variability in the loudness of the audio components of the experiment. A sound level meter was employed to ensure that maximum sound pressure levels were kept at safe levels and that the various audio components of the tasks were properly balanced. Medical screening devices were used to confirm that participants met the minimum hearing and vision standards.

a. Computer Equipment

The experimental tests were performed on a *Dell XPS* desktop computer operating with *Windows Vista*. The computer was equipped with a *NVIDIA GeForce 7800 GTX* video card and a *Realtek AC'97 Audio* sound card. A wireless keyboard and mouse were employed as input devices. The output device was a 60 cm *Dell 2405FPW* flat screen monitor with a resolution of 1920

by 1200 pixels, a 32-bit color setting and a refresh rate of 59 Hz. The monitor was located approximately 75 cm from the participants' faces.

b. Eye Tracker

Eye tracking was achieved through the use of a *Seeing Machines* camera system. The two cameras were equipped with 12 mm lenses fitted with infrared filters and were located 5 cm below the bottom edge of the monitor, along with an infrared light source. The eye tracker's cameras were connected to an HP laptop computer running *faceLab*'s (version 5.0) eye tracking software.

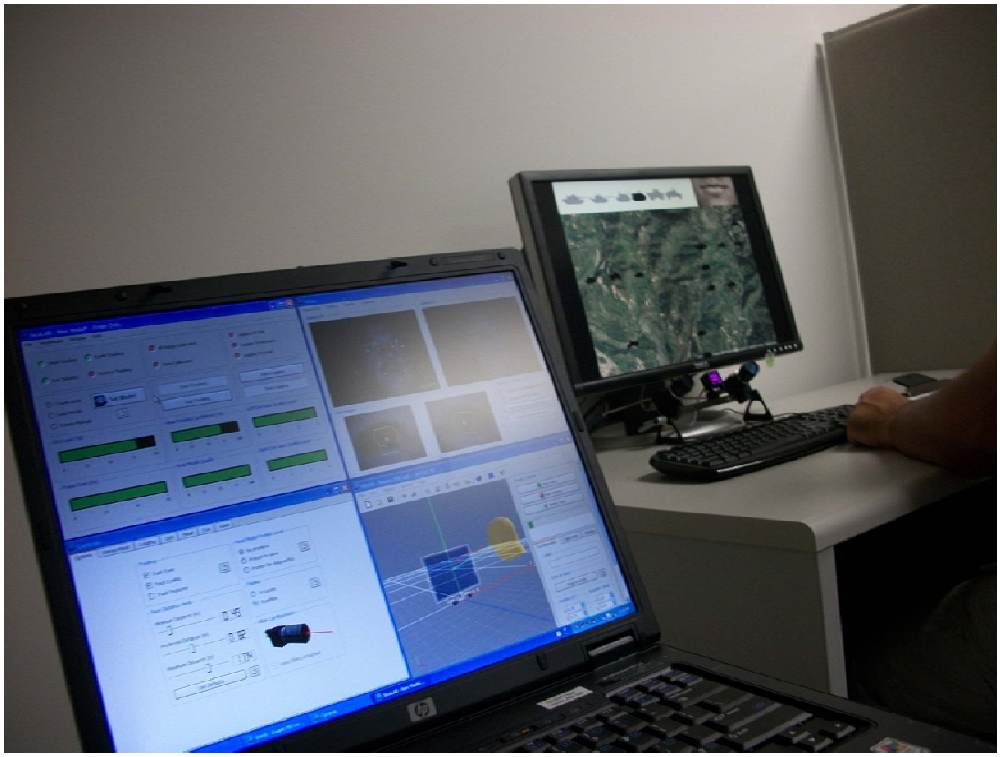

Figure 7. Eye-tracking software measuring a participant's gaze during testing session

Gaze data were collected during the experimental tests, with the researcher annotating the beginning of each of the 96 tasks the individual participants completed. The log files generated were converted to "space

separated" text files. Microsoft Excel 2007 was used to determine the number of frames in which a participant's gaze dwelled on the facial avatar.

c. Headphones

To minimize the influence of ambient sounds and reduce the variability of the loudness of the auditory portions of the tasks, the participants wore *Altec Lansing AHP 524* headphones for the duration of the experimental testing (approximately 20 minutes). Although the headphones had left and right ear indications on them, the auditory portions of the tasks were presented as monophonic (vice stereophonic) sounds; therefore participants were instructed to disregard the left-right orientation of the headphones.

d. Sound Level Meter

Sound pressure levels were measured with a *General Radio Company Permissible Sound Level Meter Type 1565-B*. A-weighted, slow averaging sound level measurements were used to set the loudness of the white noise (71 dB), auditory tasks (70 dB) and verbal messages (62 dB). Due to the logarithmic nature of the decibel scale, the sound pressure level of the combined audio signals never exceeded 74 dB, well below the 85 dB threshold for potential hearing damage.

e. Eye Chart and Audiometer

Visual acuity was tested using a wall mounted *Graham-Field No. 1240* eye chart, with a viewing distance of twenty feet. Auditory acuity was tested using a *Beltone Model 110* audiometer. The white noise generating feature of the audiometer was used to crosscheck the operation of the sound pressure meter. (Sound pressure level equals hearing level plus twenty decibels, for white noise; SPL = HL + 20 dB.)

D. RESEARCH DESIGN

The research design was a 2 x 2 x 2 x 2 factorial design. There were four independent variables with two levels each: sensory input modality of speech (auditory/visual and auditory-only); spoken sentence difficulty (high and low predictability); task type (visual-spatial and auditory); and, task difficulty (high and low). The experimental design matrix is displayed in Table 1.

There were three dependent variables: task performance, speech perception in noise and gaze dwell time. Each participant performed each of the 16 conditions six times, for a total of 96 tasks.

1. Independent Variables

a. Speech Modality

Speech modality refers to the manner in which the verbal task was presented. The spoken sentences were presented in either an auditory-visual format (with a facial avatar) or an auditory-only format (without a facial avatar). This was the main variable of interest.

b. Sentence Predictability

SPIN sentence lists provide a balance of high and low sentence predictabilities. High predictability sentences are structured so that the sentence provides contextual clues to identity of the last word of the sentence (the target word). Low predictability sentences are structured so that the sentence does not provide any indication of the last word of the sentence. The high and low predictability sentences are equivalent to the speech modality having low and high difficulty levels, respectively.

c. Task Type

Because military tasks rarely involve just a single sensory modality, it was important to expose the participants to both auditory and visual tasks.

The goal of the visual tasks was for the participant to count, or estimate, the number of target icons presented on the screen. The potential targets were presented across the top of the screen. The distracter icons were colored grey while the target icon was colored black. The icons to be scanned were presented on the lower portion of the screen. The number and placement of the icons to be searched were randomly generated.

The goal of the auditory tasks was for the participant to identify whether the change was “up” or “down” (i.e., the frequency shifted higher or lower). This auditory task was based on the JOCRF Pitch Discrimination Test (Acton and Schroeder, 2001).

d. Task Difficulty

Each of the two types of concurrent tasks (auditory and visual) was presented at two difficulty levels. The purpose of exposing the participants to two difficulty levels was to attempt to determine if the efficacy of the facial avatar was related to the difficulty of the concurrent tasks.

Low difficulty visual tasks consisted of four potential target icons; all oriented the same direction (i.e., facing right). High difficulty tasks consisted of six potential target icons randomly oriented (i.e., randomly facing either right or left). In either case, the participants were limited to five seconds to determine the number of target icons.

The level of difficulty for the auditory tasks was related to the degree to which the test tone changed. The initial tone was always presented at 435 Hz; for the low difficulty auditory tasks the second tone was either 425 or 445 Hz (a difference of 10 Hz), for the high difficulty tasks the second tone was either 430 or 440 Hz (a difference of only 5 Hz).

2. Dependent Variables

a. Word Identification

The effectiveness of the facial avatar was primarily measured via the correct identification of the target word at the end of each sentence. Incorrect spelling of the correct target word did not count as an error. Word identification score was measured as a percentage and determined by dividing the number of correct responses by the total number of exposures for the given combination of independent variables. As a 2 x 2 x 2 x 2 research design, there were 16 different combinations and six exposures to each combination, yielding a total of 96 tasks.

b. Task Performance

Performance on the concurrent tasks was intended to provide insight into the "cognitive cost" of presenting a facial avatar. Auditory task performance was scored as the percentage of correct participant observations of whether the tone "went up" or "went down." Visual task performance was scored as the percentage of correct participant observations of the number of target icons presented on the screen. As a 2 x 2 x 2 x 2 research design, there were 16 different combinations and six exposures to each combination, yielding a total of 96 tasks.

c. Gaze Dwell Time

Determination of the amount of time the participants' gaze dwelled in the area occupied by the facial avatar was intended to provide insight into the degree to which the participants attended to the facial avatar. The period of interest began when the verbal message started (one second after the task began) and lasted for three seconds. At an eye tracking capture rate of 60 frames per second, the maximum number of frames in which the participant focused his visual attention on the area of the facial avatar was 180. Gaze dwell

time was scored as a percentage of the period of interest that the participants gazed in the area normally occupied by the facial avatar.

3. Test Design

The verbal, visual and auditory task components were combined to produce the experimental tests using *Microsoft PowerPoint 2007*. The task components were embedded into the slides; the video and audio files were set to begin automatically when the participants advanced the slideshow to a task slide. Each task ended automatically after five seconds and the slideshow advanced to a slide instructing the participants to record their observations. The participants were given as much time as they needed to record their observations, the next task began when the participants advanced the slideshow again. Every task exposure consisted of a verbal task and either a visual task or an auditory task. Each of the sixteen combinations of the four independent variables was equally represented. The order of the task combinations was arranged randomly. All participants received the same random arrangement.

		Auditory-Visual Speech Modality		**Auditory-Only Speech Modality**	
		High Predictability Sentence	**Low Predictability Sentence**	**High Predictability Sentence**	**Low Predictability Sentence**
Visual Task	**High Task Difficulty**				
	Low Task Difficulty				
Auditory Task	**High Task Difficulty**				
	Low Task Difficulty				

Table 1. Research design–Matrix of independent variables

Since each target word was presented twice in each test (once in a high predictability sentence and once in a low predictability sentence), if the high predictability sentence was presented in conjunction with the facial avatar the low predictability sentence was presented in the auditory-only mode, and vice versa. Two experimental tests were developed to account for any innate differences in the comprehensibility of the verbal messages. The second iteration of the experimental test reversed the speech modality of each sentence; auditory-visual presentation of a sentence in the first experimental test was matched by an auditory-only presentation in the second experimental test.

The assignment of participants to the two variants of the experimental tests was pseudorandom; the participants were assigned to the two test variants alternately.

E. PROCEDURE

The participants performed a series of tasks on a computer while listening to spoken sentences in a noisy environment. Each task consisted of a five-second exposure to either visual or auditory stimuli, while concurrently listening to a spoken sentence (presented with or without visual cues). This was followed by a participant-controlled period of time during which the participant reported their observations. The participant then initiated the next task. The exposure to the entire series of tasks typically lasted approximately 20 minutes. The eye tracker was employed to determine the measure participant's gaze while performing the tasks.

1. Consent

Before beginning their involvement in the study, participants read and signed a voluntary consent form, including consent for audio-video recording. Although the eye tracking system does not record sound or images, it does display the participant’s image temporarily on a connected laptop computer and does record the participant’s head and eye movements.

2. Screening

Participants underwent a brief visual and auditory acuity screening process to ensure they met the minimum vision and hearing standards.

3. Eye Tracker Calibration

Participants who did not wear glasses took part in the collection of eye tracking data. To maximize the accuracy of the eye tracking, the system was calibrated to the participants' features and their gaze was calibrated using a series of nine marked screen positions.

4. Training

Participants underwent a brief training session to familiarize themselves with the verbal, visual and auditory tasks. The training was performed using *Microsoft PowerPoint 2007*, and gradually exposed the participants to each of the three task components.

During the training period, participants were informed that the auditory and visual tasks were their primary task; the verbal task (identification of the spoken target word) was the task of secondary importance. The rationale for assigning priority to the concurrent tasks rather than the verbal task was that the effect of the facial avatar on concurrent tasks was an important research question. Additionally, all the participants needed to divide their cognitive resources in a similar manner.

The training session lasted approximately ten minutes and participants were allowed to repeat portions of training session if they chose to do so.

5. Testing

The experimental testing immediately followed the training session. Participants completed 96 experimental tasks; each task combined a verbal message with either a visual or auditory task. Participants were permitted to

proceed through the experimental tasks at their own pace; as with the training session, the participant initiated each task exposure and were allowed as much time as necessary between tasks to record their observations. The experimental testing session lasted approximately twenty minutes.

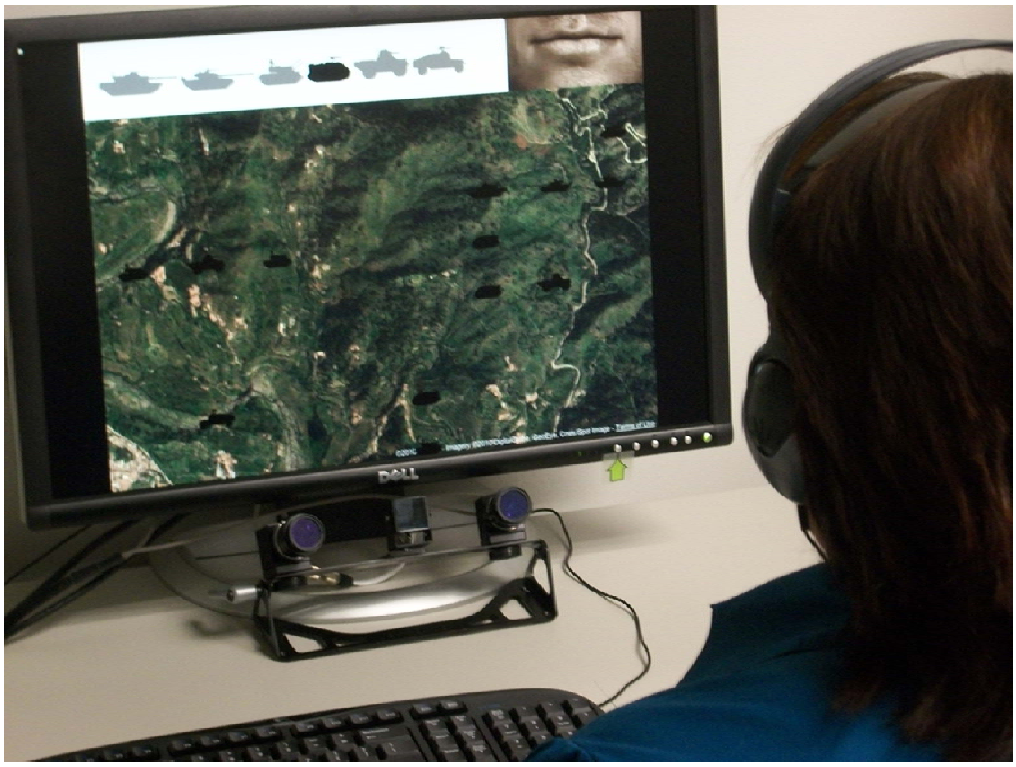

Figure 8. A visual task with animated facial avatar (note the eye tracking cameras below the monitor)

During the experimental testing session, the investigator monitored the participant's progress. If eye tracking was employed, the researcher annotated the beginning of each task exposure in the eye tracking log file.

6. Debrief

After completion of the testing session, the investigator informed the participants that the purpose of the investigation was to assess the extent to which the facial avatar improves performance on a task while performing a concurrent task. The participants were given the opportunity to ask questions

regarding the study, and were requested not to divulge the nature of the research to other Naval Postgraduate School students until after the data collection period had concluded.

Participants were thanked for their assistance, informed that they could be notified of the results of the study and were offered a copy of their signed consent forms.

IV. RESULTS

Twenty participants completed the experimental testing over a three-week period. After the data were collected, the data were organized using *Excel* and analyzed using *Minitab*. The three dependent variables were examined for the main effects, as well as any interactions that may have been present.

A. WORD IDENTIFICATION

1. Suitability of ANOVA

In order to perform an ANOVA, the data must be considered independent, normally distributed and homoscedastic. There was no reason to believe that any of the results were unduly influenced by the performance of previous participants. Participants were requested not to divulge any information regarding the experimental testing to any other potential participants until after the data collection phase was completed. The experimental testing was performed in the same manner with all participants and no changes were made to any of the test parameters, such as sound pressure levels or monitor screen size/position.

To determine if the data were normally distributed the Word Identification scores were examined both graphically and using the Ryan-Joiner normality test. Figure 9 indicates that the data were roughly normally distributed, but there appeared to be some skewness.

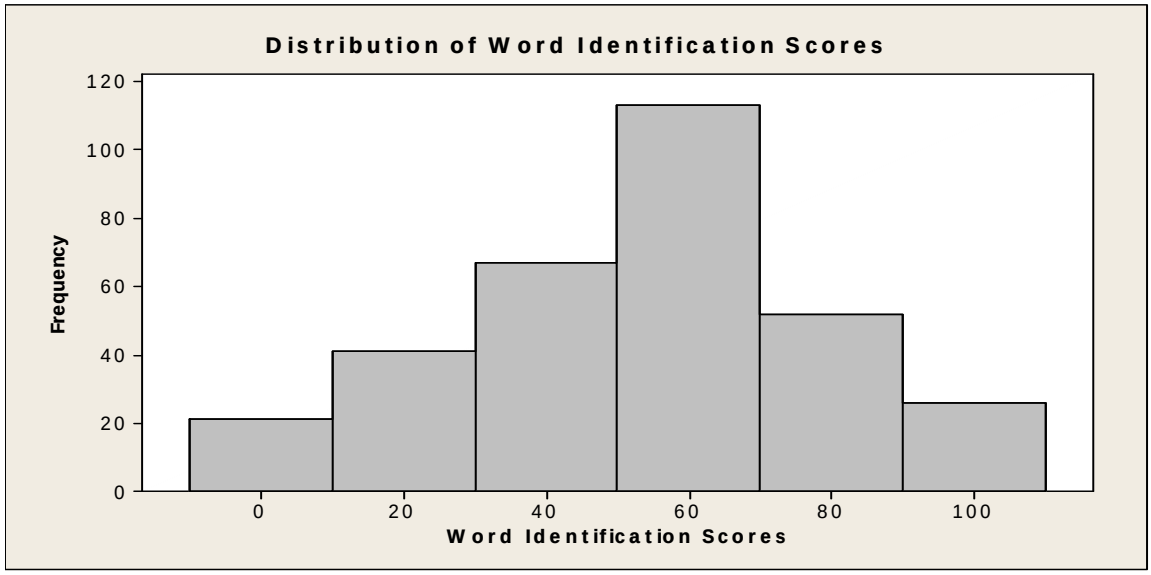

Figure 9. Distribution of Word Identification scores

However, the Ryan-Joiner normality test (Figure 10) confirmed that the data were normally distributed, with a Ryan-Joiner statistic of 0.997 and $p>0.100$ (the null hypothesis of this test is that the data are correlated with a normal distribution).

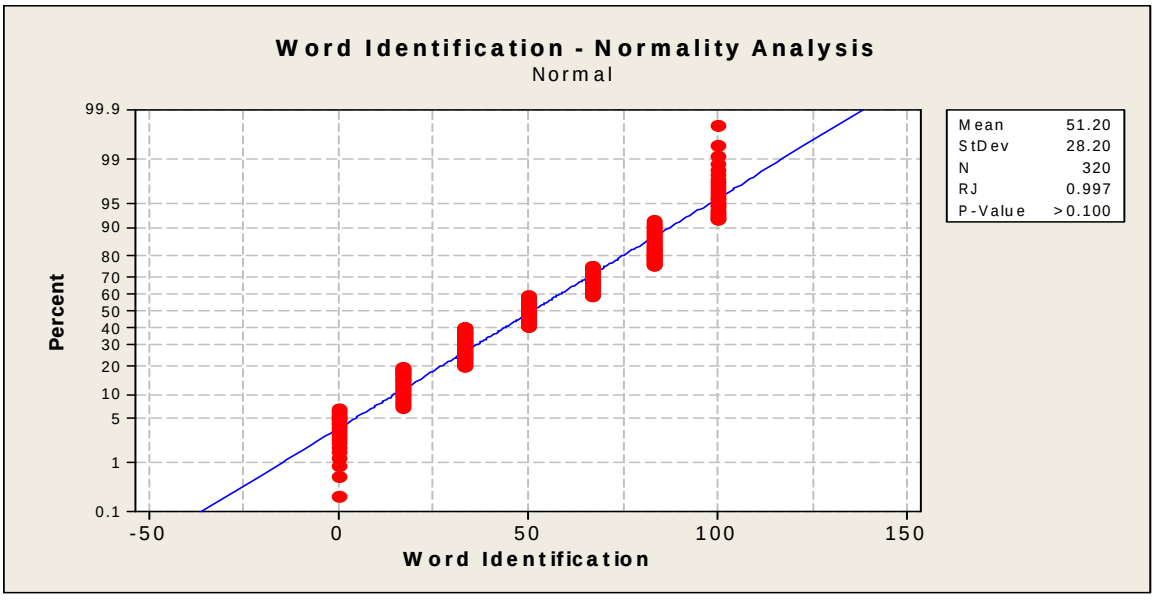

Figure 10. Ryan-Joiner normality test for Word Identification

Homoscedasticity was determined through graphical analysis of the residuals. Figure 11 indicates that the residuals are normally distributed and fall along the normal line with little deviation.

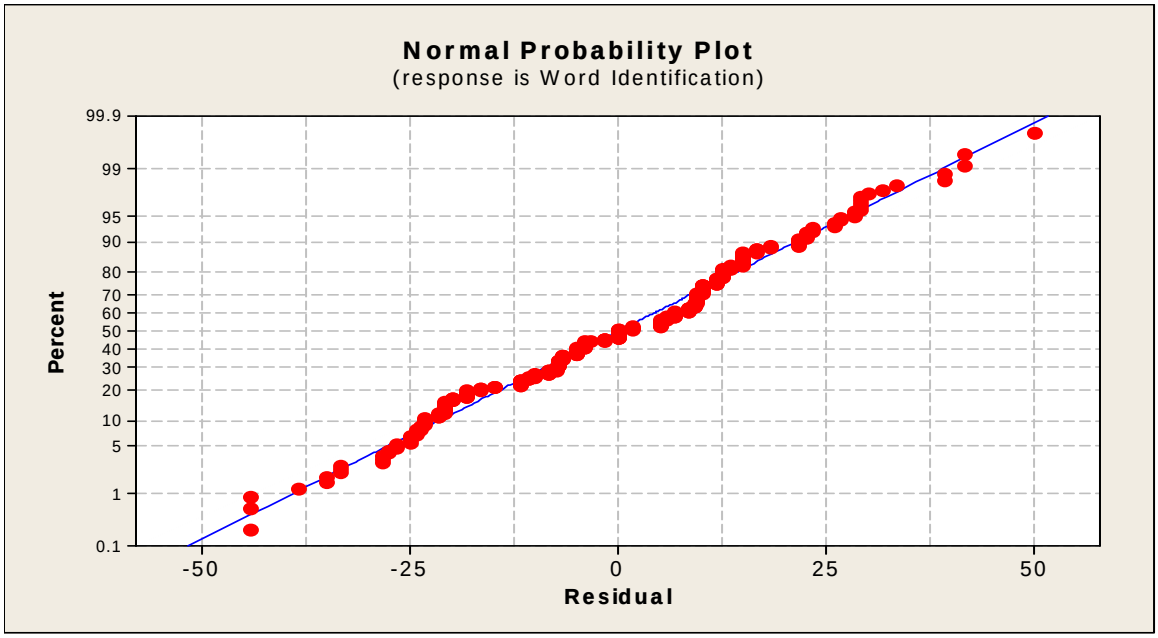

Figure 11. Normal probability plot of the residuals of the Word Identification scores

These tests indicate that the Word Identification scores are independent, normally distributed and homoscedastic. Therefore, the Word Identification scores are suitable for ANOVA data analysis.

2. Overall Results

The overall test results for the Word Identification task are presented in Table 2 as means of the scores of the twenty participants for each combination of the independent variables at their two levels.

Word Identification Scores		Auditory-Visual Speech Modality		Auditory-Only Speech Modality	
		High Sentence Predictability	Low Sentence Predictability	High Sentence Predictability	Low Sentence Predictability
Auditory Task	High Task Difficulty	71.7 (15.9)	31.7 (12.3)	70.8 (16.1)	24.2 (16.6)
	Low Task Difficulty	76.7 (15.7)	41.7 (19.1)	66.7 (7.9)	28.3 (15.3)
Visual Task	High Task Difficulty	60.8 (26.6)	35.0 (18.2)	54.2 (18.6)	20.0 (18.9)
	Low Task Difficulty	73.3 (11.7)	23.3 (16.6)	90.6 (11.3)	33.3 (20.9)

Table 2. Mean Word Identification scores (standard deviation in parentheses)

Although there appeared to be vast differences between the individual cells, any significant differences needed to be revealed through the use of ANOVA. Table 3 displays the results of the ANOVA statistical analysis.

Word Identification	DF	Seq SS	Adj SS	Adj MS	F	P
Mode	1	147	730	730	2.49	0.116
Predictability	***1***	***138195***	***114723***	***114723***	***390.93***	***0.000***
Task Type	1	43	478	478	1.63	0.203
Task Difficulty	***1***	***6398***	***4604***	***4604***	***15.69***	***0.000***
Mode x Predictability	1	1854	700	700	2.38	0.124
Mode x Task Type	***1***	***1811***	***1484***	***1484***	***5.06***	***0.025***
Mode x Task Difficulty	***1***	***3043***	***1230***	***1230***	***4.19***	***0.041***
Predictability x Task Type	1	0	56	56	0.19	0.663
Predictability x Task Difficulty	***1***	***3287***	***1230***	***1230***	***4.19***	***0.041***
Task Type x Task Difficulty	***1***	***1371***	***1354***	***1354***	***4.62***	***0.032***
Mode x Predictability x Task Type	1	16	33	33	0.11	0.737
Mode x Predictability x Task Difficulty	1	21	21	21	0.07	0.788
Mode x Task Type x Task Difficulty	***1***	***4373***	***4373***	***4373***	***14.90***	***0.000***
Predictability x Task Type x Task Difficulty	***1***	***3929***	***3929***	***3929***	***13.39***	***0.000***
Mode x Predictability x Task Type x Task Difficulty	1	5	5	5	0.02	0.893
Error	304	89213	89213	293		
Total	319	253707				

Table 3. Results of ANOVA of Word Identification scores (significant results are in bold italics)

3. Main Effects

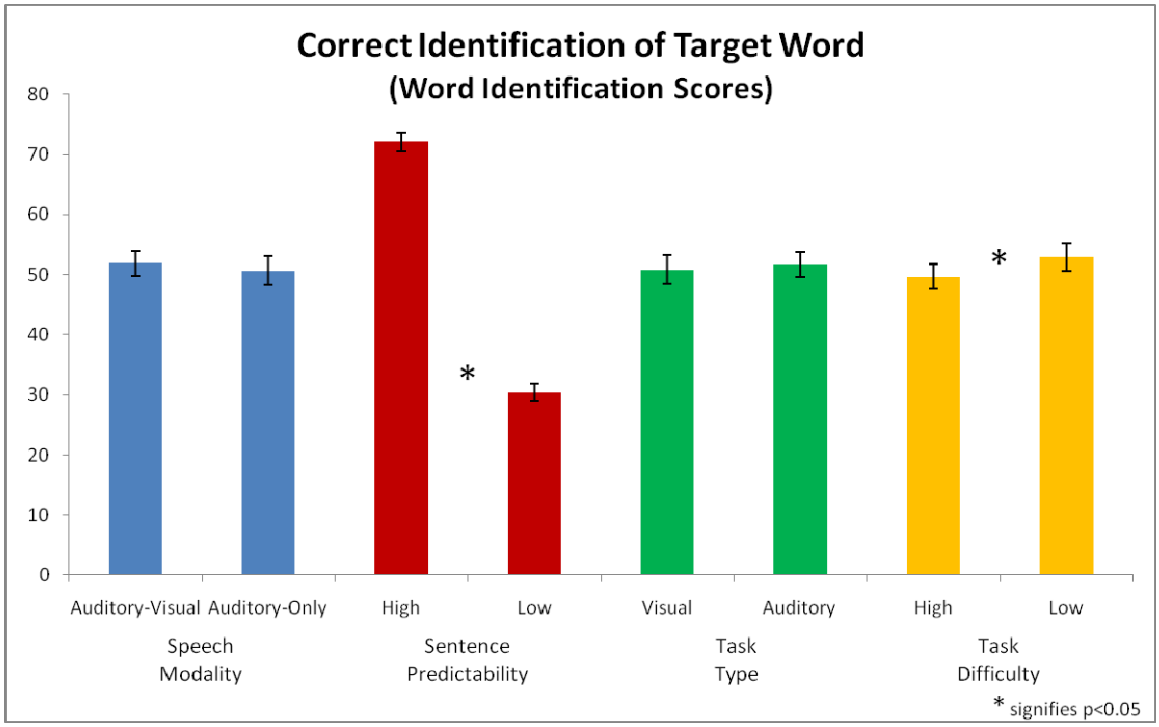

Figure 12. Word Identification–Main effects between the levels of the four independent variables, means with standard error bars (***** indicates significant difference)

For Word Identification, the only independent variables that resulted in significant differences between their levels were Sentence Predictability and Task Difficulty. Figure 12 displays the small but significant difference between the two levels of Task Difficulty, $F(1,304)=15.7$, $p<0.001$; and the large and significant difference between the two levels of Sentence Predictability, $F(1,304)=390.9$, $p<0.001$.

Surprisingly, there was no significant difference between the Auditory-Visual and Auditory-Only levels of Speech Modality, $F(1,304)=2.49$, $p=0.116$.

4. Interactions

For Word Identification, several interactions were found to be significant: Speech Modality were Speech Modality by Task Type $F(1,304)=5.06$, $p=0.025$; Speech Modality by Task Difficulty $F(1,304)=4.19$, $p=0.041$; and, Speech Modality by Task Type by Task Difficulty $F(1,304)=14.9$, $p<0.001$. These interactions had the potential to provide insight into the efficacy of the facial avatar and were examined further.

The significant interactions not involving Speech Modality were: Sentence Predictability by Task Difficulty $F(1,304)=4.19$, $p=0.041$; Task Type by Task Difficulty $F(1,304)=4.62$, $p=0.032$; and, Sentence Predictability by Task Type by Task Difficulty, $F(1,304)=13.4$, $p<0.001$. Although these interactions were significant, they did not help support or oppose the efficacy of the facial avatar. Therefore, no further analyses of these interactions were performed.

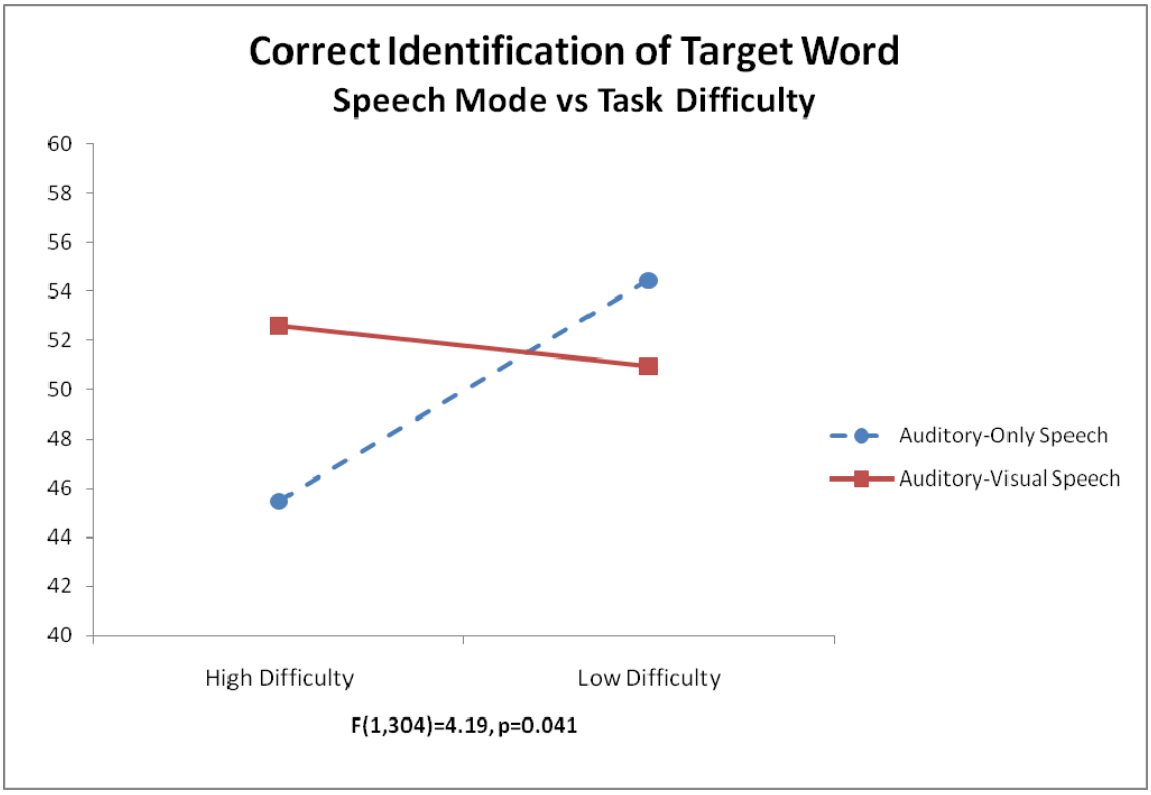

Figure 13. Interaction between Speech Modality and concurrent task difficulty (Word Identification scores)

Figure 13 demonstrates the interaction between Speech Modality and Task Difficulty. Participants performed 7.1 percentage points better at identifying the target word with the presence of the facial avatar (52.6 with vice 45.5 without) at the higher task difficulty. At the lower task difficulty, when the facial avatar was present, performance was 3.4 percentage points worse performance at identifying the target word (51.0 with the avatar vice 54.4 without the avatar).

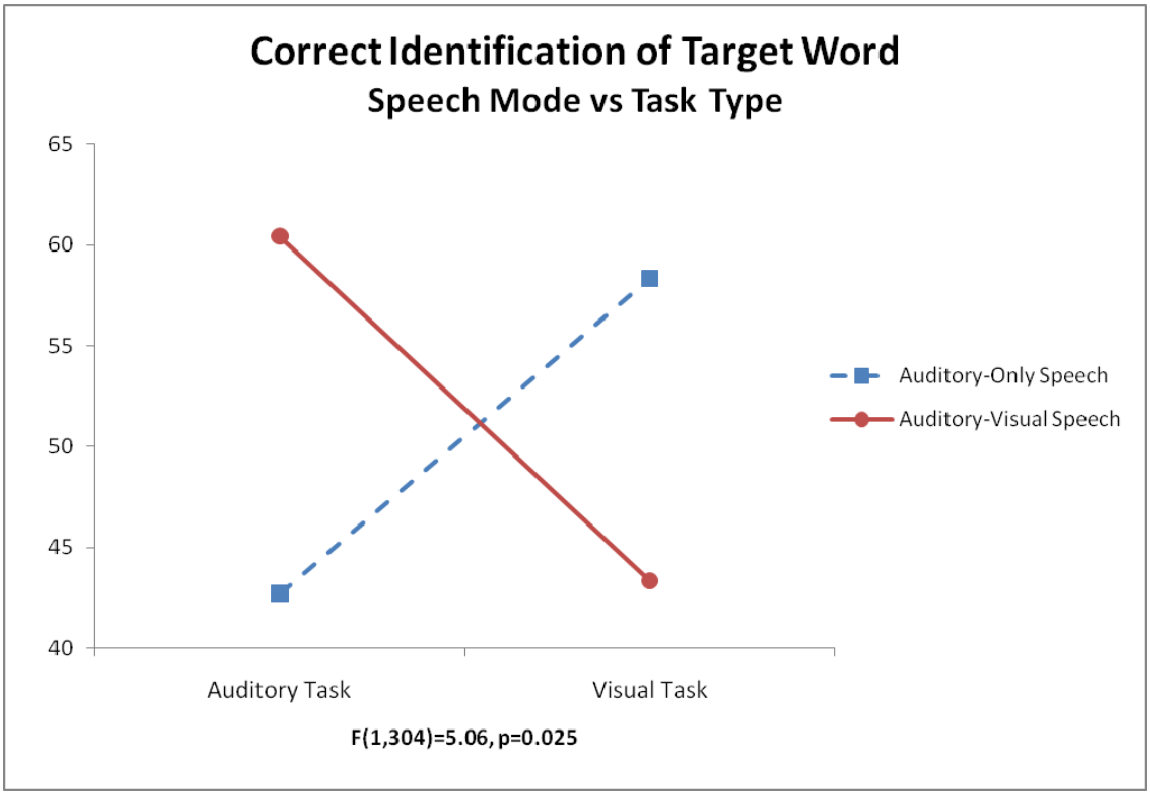

Figure 14. Interaction between Speech Modality and concurrent task type (Word Identification scores)

Figure 14 demonstrates the interaction between Speech Modality and Task Type. Participants performed 17.7 percentage points better at identifying the target word with the presence of the facial avatar (60.4 with the avatar vice 42.7 without the avatar) during auditory tasks. During visual tasks, the presence of the facial avatar coincided with 15.0 percentage point worse performance at identifying the target word (43.3 with the avatar vice 58.3 without the avatar).

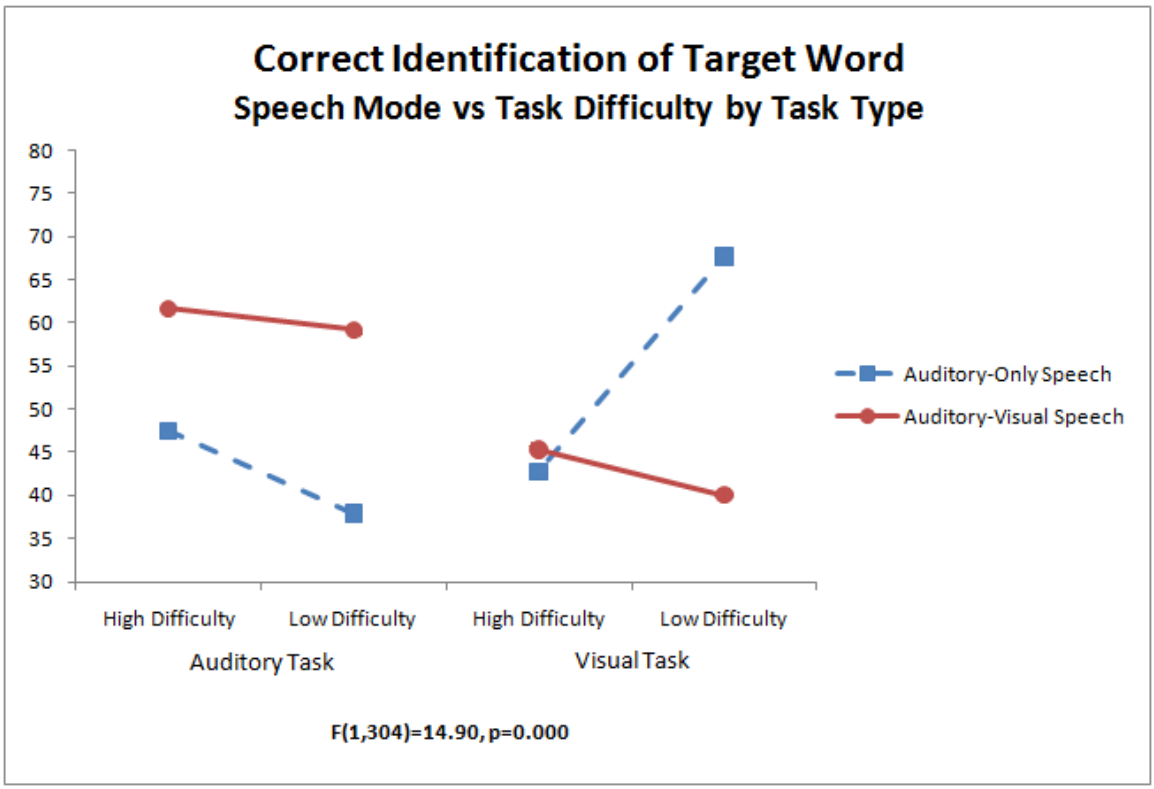

Figure 15. Interaction between Speech Modality and Task Difficulty and Task Type (Word Identification scores)

Figure 15 demonstrates the interaction between Speech Modality, Task Difficulty and Task Type. While performing auditory tasks, participants performed consistently better at identifying the target word when the facial avatar was present (high difficulty: 61.7 with vice 47.5 without; low difficulty: 59.2 with vice 37.9 without). However, while performing visual tasks, participants performed roughly equally during high difficulty tasks but performed much better during low difficulty visual tasks when the facial avatar was not present (high difficulty: 45.3 with vice 42.8 without; low difficulty: 40.0 with vice 67.7 without). The presence of the facial avatar corresponds to a consistent improvement in Word Identification scores when the concurrent task is an auditory task, to equal performance with high difficulty visual tasks and a decrease in performance during low difficulty visual tasks.

B. TASK PERFORMANCE

1. Suitability of ANOVA

Task Performance scores were defined as the percentage of correct responses for the concurrent auditory and visual tasks. Task Performance data had to be independent, normally distributed and homoscedastic in order to be analyzed using ANOVA. Similar to the Word Identification data, there was no reason to believe that any of the results relating to the performance of concurrent tasks were unduly influenced by the performance of previous participants.

To determine if the data were normally distributed, the Task Performance scores were examined both graphically and using the Ryan-Joiner normality test. Figure 16 indicates that the data were skewed and that a ceiling effect was present. A more normally distributed set of data may have been generated if the concurrent tasks were more difficult. Considering the existing data were composed of two difficulty levels and two task types, the data were distributed relatively normally; however, an objective test of normality needed to be performed. The Ryan-Joiner normality test (Figure 17) confirmed that the data were normally distributed, with a Ryan-Joiner statistic of 0.998 and $p>0.100$.

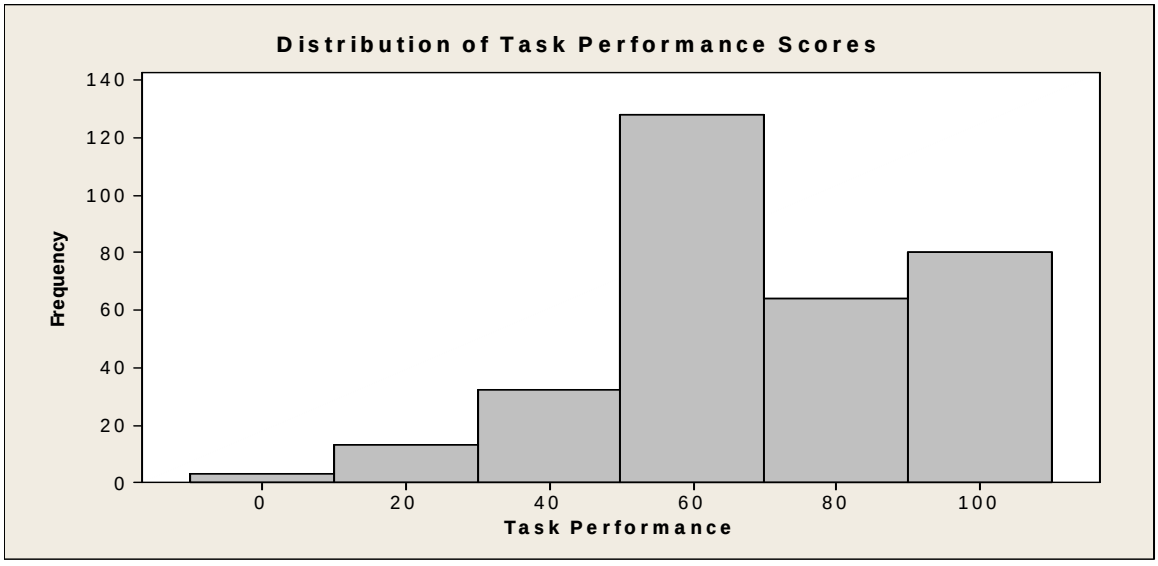

Figure 16. Distribution of Task Performance scores

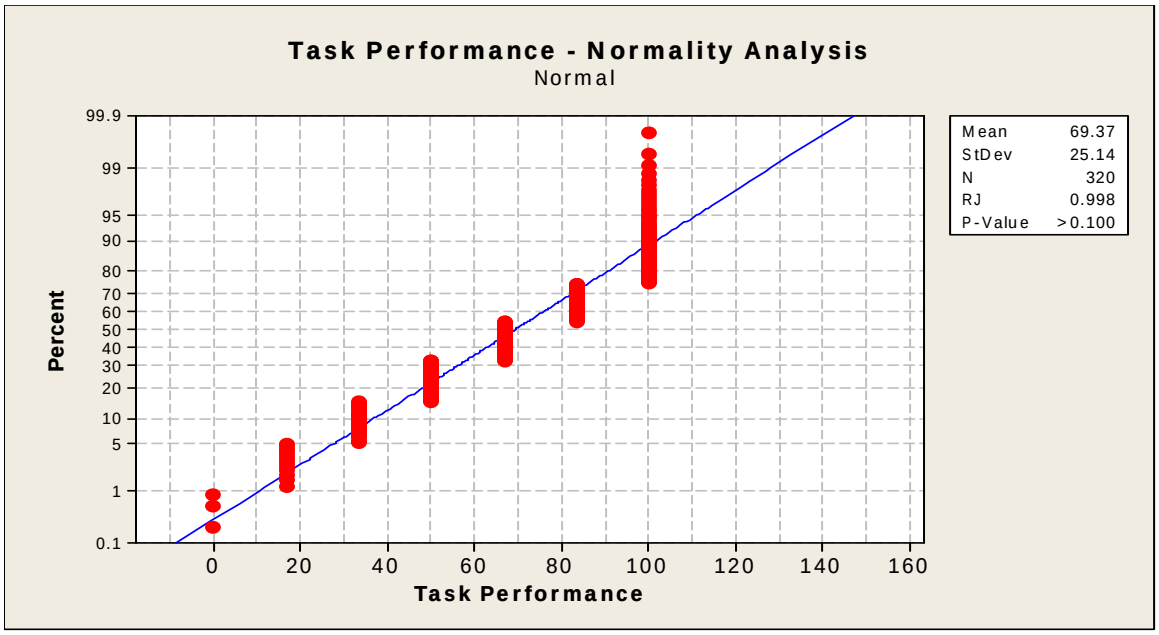

Figure 17. Ryan-Joiner normality test for Task Performance

Homoscedasticity was determined through graphical analysis of the residuals. Figure 18 indicates that the residuals are normally distributed and fall along the normal line with little deviation, although the ceiling effect is apparent.

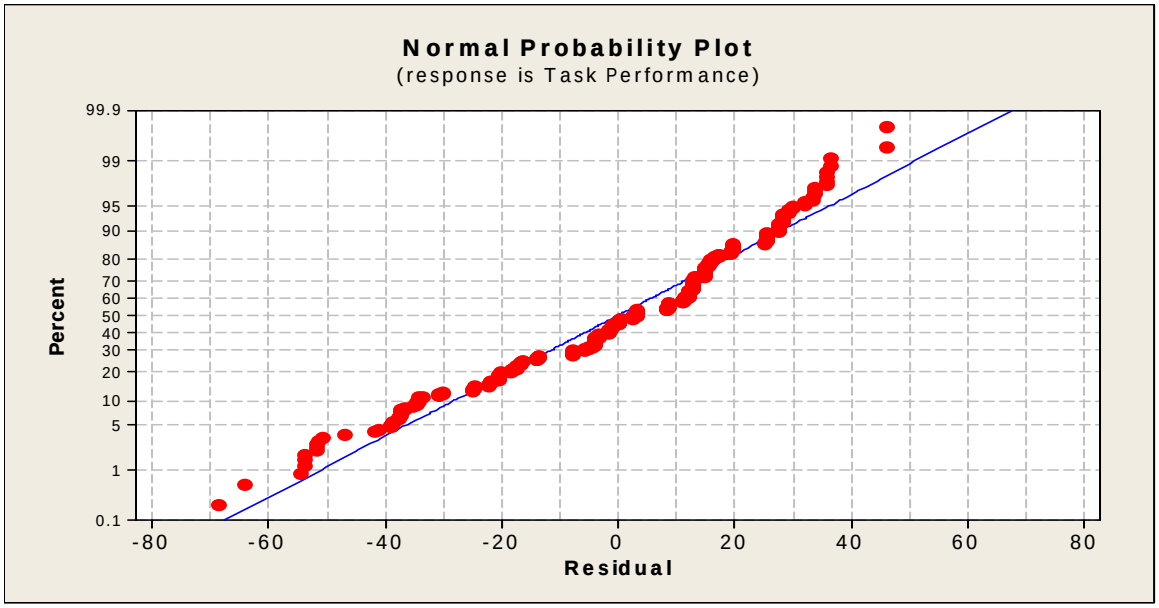

Figure 18. Normal probability plot of the residuals of the Task Performance scores

These tests indicate that the Task Performance scores are independent, normally distributed and homoscedastic. The Task Performance scores are suitable for ANOVA data analysis.

2. Overall Results

The overall test results for Task Performance are presented in Table 4 as means of the scores of the twenty participants for each combination of the independent variables at their two levels.

Task Performance Scores		Auditory-Visual Speech Modality		Auditory-Only Speech Modality	
		High Sentence Predictability	Low Sentence Predictability	High Sentence Predictability	Low Sentence Predictability
Auditory Task	High Task Difficulty	78.9 (18.5)	71.7 (23.6)	76.7 (23.2)	76.7 (23.2)
	Low Task Difficulty	79.2 (22.9)	89.2 (20.4)	80.0 (24.6)	82.2 (24.7)
Visual Task	High Task Difficulty	51.7 (18.7)	52.8 (24.0)	40.8 (20.6)	51.7 (24.2)
	Low Task Difficulty	78.3 (19.3)	66.7 (17.1)	67.2 (18.8)	66.7 (26.5)

Table 4. Mean Task Performance scores (standard deviation in parentheses)

There appeared to be differences between the individual cells, but any significant differences needed to be revealed through the use of ANOVA. Table 5 displays the results of the ANOVA statistical analysis.

Task Performance	DF	Seq SS	Adj SS	Adj MS	F	P
Mode	1	86.8	746.1	746.1	1.56	0.213
Predictability	1	347.2	23.9	23.9	0.05	0.823
Task Type	***1***	***35420.1***	***26954.4***	***26954.4***	***56.27***	***0.000***
Task Difficulty	***1***	***12635.8***	***12639.0***	***12639.0***	***26.39***	***0.000***
Mode x Predictability	1	122.5	440.6	440.6	0.92	0.338
Mode x Task Type	1	99.1	416.7	416.7	0.87	0.352
Mode x Task Difficulty	1	7.0	69.5	69.5	0.15	0.703
Predictability x Task Type	1	8.4	29.8	29.8	0.06	0.803
Predictability x Task Difficulty	1	21.7	23.9	23.9	0.05	0.823
Task Type x Task Difficulty	***1***	***4084.1***	***3273.9***	***3273.9***	***6.83***	***0.009***
Mode x Predictability x Task Type	1	467.9	490.2	490.2	1.02	0.313
Mode x Predictability x Task Difficulty	1	198.5	198.5	198.5	0.41	0.520
Mode x Task Type x Task Difficulty	1	101.3	101.3	101.3	0.21	0.646
Predictability x Task Type x Task Difficulty	***1***	***2037.8***	***2037.8***	***2037.8***	***4.25***	***0.040***
Mode x Predictability x Task Type x Task Difficulty	1	287.8	287.8	287.8	0.60	0.439
Error						
Total	304	145615.7	145615.7	479.0		

Table 5. Results of ANOVA of Task Performance scores (significant results are in bold italics)

3. Main Effects

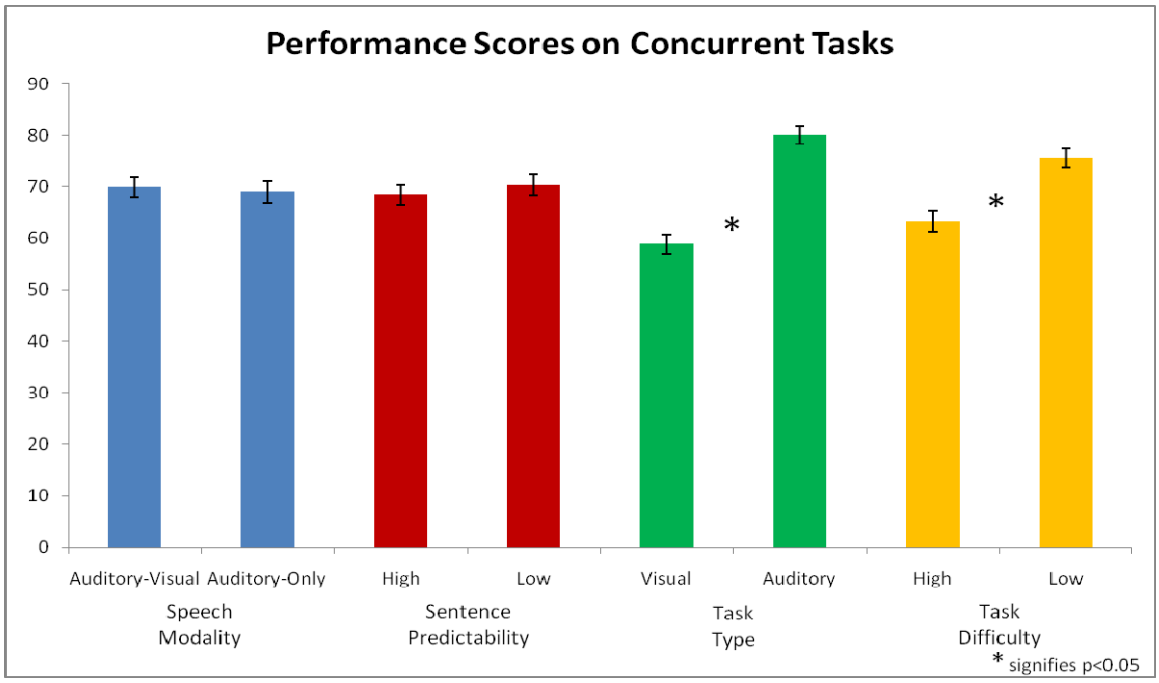

Figure 19. Task Performance—Main effects between the levels of the four Independent Variables, means with standard error bars (***** indicates significant difference)

For Task Performance, the two independent variables that possessed significant differences between their levels were Task Type and Task Difficulty. Figure 19 displays the significant difference between the two levels of Task Type, $F(1,304)=56.3$, $p<0.001$, and two levels of Task Difficulty, $F(1,304)=26.4$, $p<0.001$.

4. Interactions

For Task Performance, two interactions were found to be significant: Task Type by Task Difficulty, $F(1,304)=6.83$, $p=0.009$; and, Sentence Predictability by Task Type by Task Difficulty $F(1,304)=4.25$, $p=0.040$. No significant interactions were found for Speech Modality (i.e., the presence of the facial avatar) had no significant interaction effects.

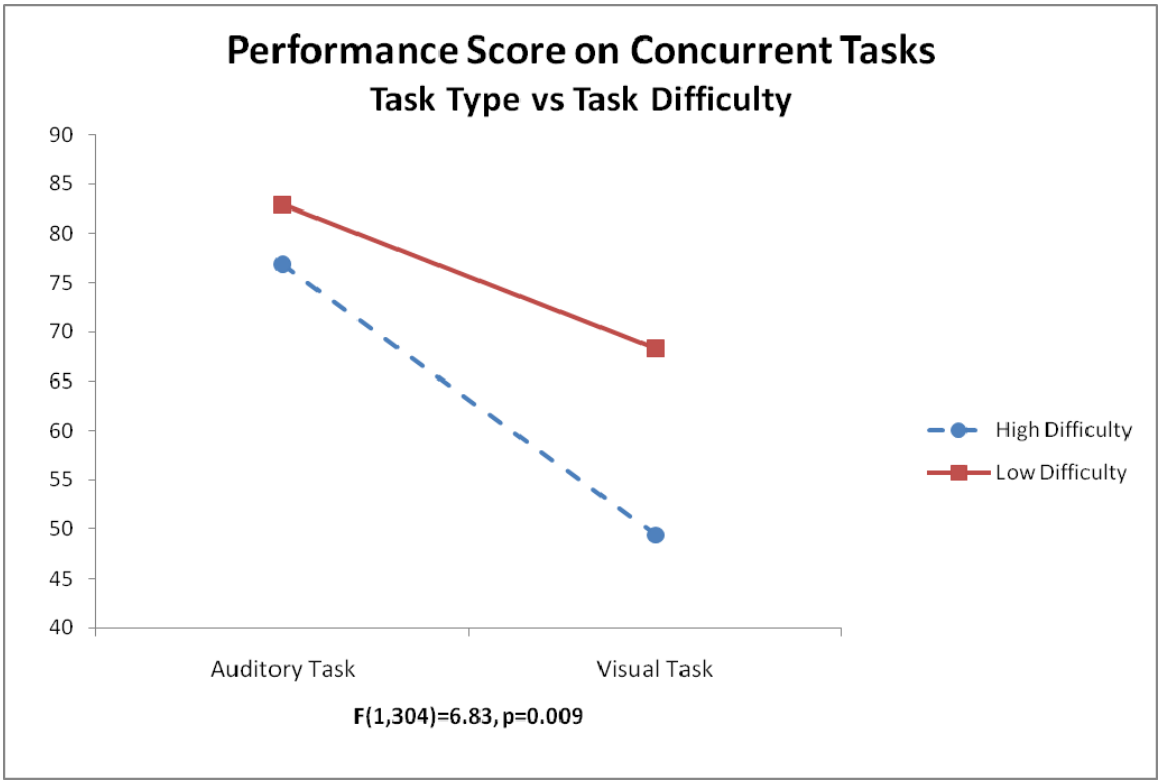

Figure 20. Interaction between Task Type and Task Difficulty (Task Performance scores)

Figure 20 demonstrates the interaction between Task Type and Task Difficulty. High difficulty tasks were indeed more difficult for both visual and auditory tasks, but also indicated that the difficulty differences between the auditory tasks were to a lesser extent that the visual tasks.

C. GAZE DWELL TIME

1. Suitability of ANOVA

The Gaze Dwell Time data also needed to be considered independent, normally distributed and homoscedastic in order to be analyzed using ANOVA. Again, there was no reason to believe that any of the results related to the participants' gaze were unduly influenced by the performance of previous participants.

To determine if the data were normally distributed the gaze dwell times were examined both graphically and using the Ryan-Joiner normality test. Figure 21 indicates that the data were severely skewed and that a floor effect was present. Although it was readily apparent that the data were not normally distributed, an objective test of normality was performed to confirm the subjective graphical analysis.

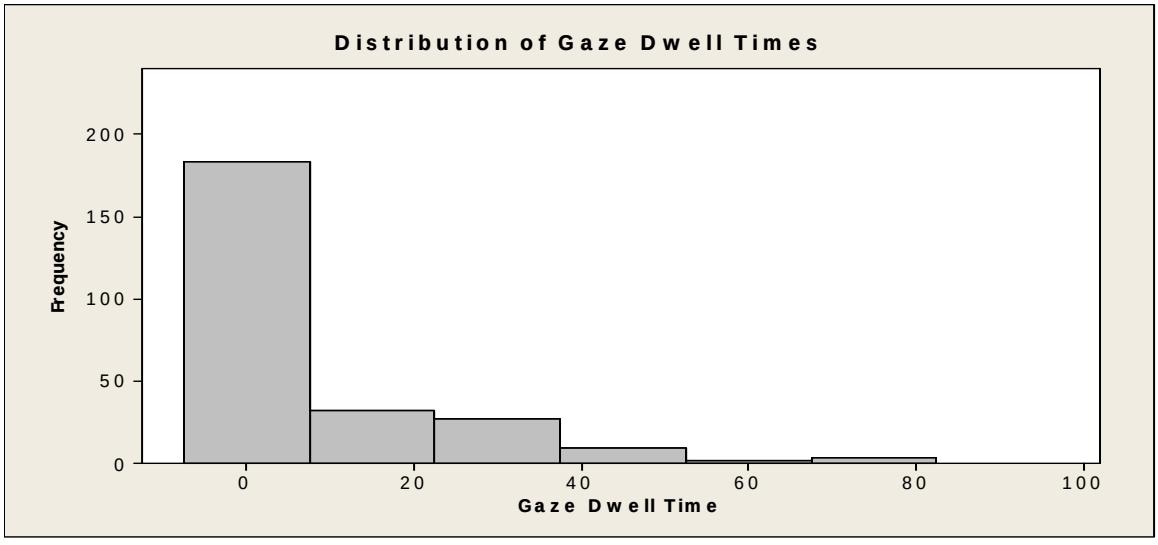

Figure 21. Distribution of Gaze Dwell Times

The Ryan-Joiner normality test (Figure 22) confirmed that the data were not normally distributed, with a Ryan-Joiner statistic of 0.854 and $p<0.010$ (i.e., the null hypothesis that the data are correlated with a normal distribution was rejected).

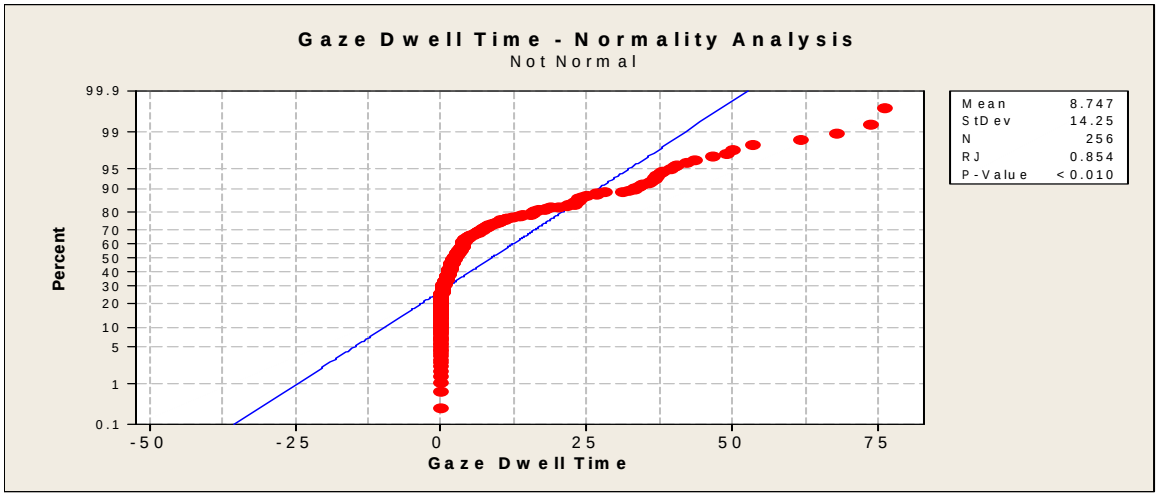

Figure 22. Ryan-Joiner normality test for Gaze Dwell Times

Homoscedasticity was determined through graphical analysis of the residuals. Figure 23 indicates that the residuals are not normally distributed and do not fall along the normal line (i.e., the data are heteroscedastic).

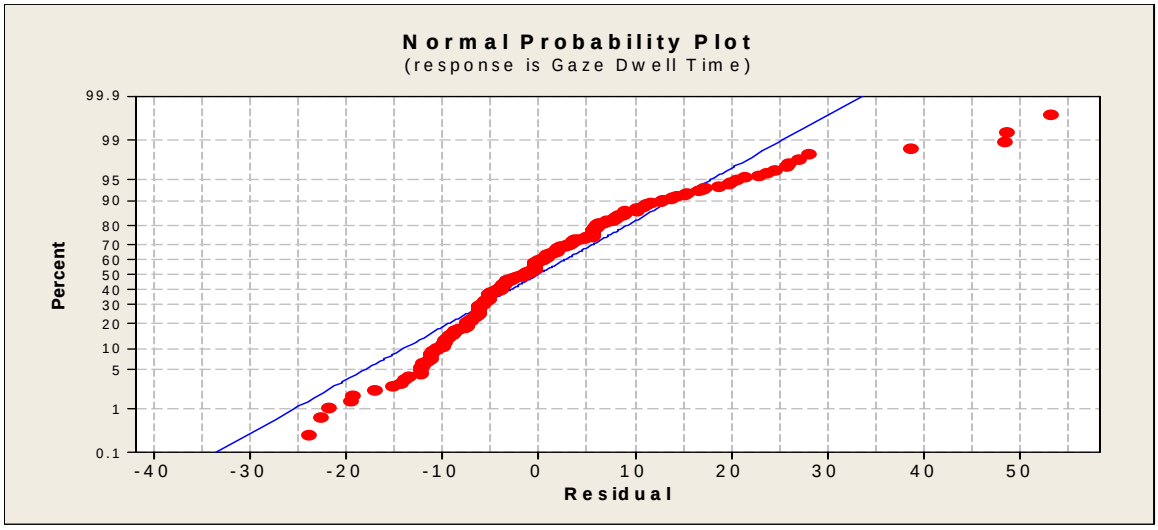

Figure 23. Normal Probability Plot of the Residuals of the Gaze Dwell Times

These tests indicate that the Gaze Dwell Times are independent, but not normally distributed and homoscedastic. The Gaze Dwell Times are unsuitable for ANOVA data analysis, but can still provide insight into the participants' eye fixations during the experimental testing. These data require analysis with non-parametric statistical tests.

2. General Observations

The overall test results for the Gaze Dwell Times are presented in Table 6 as means of the scores of the sixteen participants that provided eye-tracking data, for each combination of the independent variables at their two levels.

Gaze Dwell Times		Auditory-Visual Speech Modality		Auditory-Only Speech Modality	
		High Sentence Predictability	Low Sentence Predictability	High Sentence Predictability	Low Sentence Predictability
Auditory Task	High Task Difficulty	29.8 (17.1)	14.7 (12.2)	3.4 (4.0)	1.8 (2.2)
	Low Task Difficulty	28.3 (19.4)	26.7 (17.2)	1.4 (2.1)	2.3 (3.4)
Visual Task	High Task Difficulty	3.9 (6.6)	3.8 (7.9)	2.9 (5.3)	1.1 (1.4)
	Low Task Difficulty	3.7 (3.0)	1.6 (2.2)	3.3 (4.5)	1.0 (1.5)

Table 6. Mean Gaze Dwell Times (standard deviation in parentheses)

There appeared to be differences between the individual cells; but due to the nature of the data, any significant differences could not be revealed through the use of ANOVA. However, Kruskal-Wallis analyses could be performed to reveal the main effects.

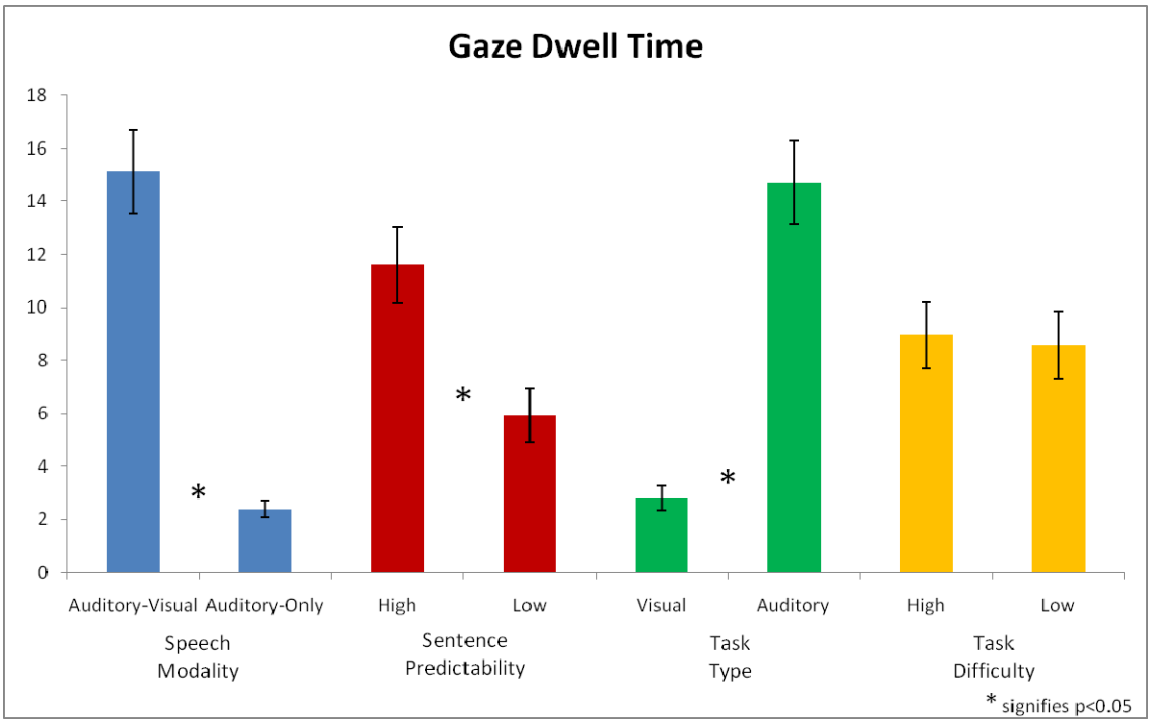

Figure 24. Gaze Dwell Times–Main effects between the levels of the four independent variables, means with standard error bars (***** indicates significant difference)

Figure 24 indicates that participants gazed directly at the area of the screen in which the facial avatar appeared for significantly longer periods of time during auditory concurrent tasks when the facial avatar was present, $H(1)=42.98$, $p<0.001$. It also appeared that participants spent significantly more time with their gaze directed to the facial avatar when the verbal messages provided contextual clues to the target word, $H(1)=14.77$, $p<0.001$ (i.e., when Sentence Predictability was high participants tended to gaze at the facial avatar for a longer duration of time). Regarding Task Type, participants had a significantly longer dwell time for auditory tasks, $H(1)=34.11$, $p<0.001$. Only Task Difficulty had no influence on the duration of time the participants gazed at the facial avatar, $H(1)=0.30$, $p<0.578$.

The mean Gaze Dwell Times were plotted in terms of Speech Mode by Task Type to produce Figure 25, displaying the relationship between the two variables.

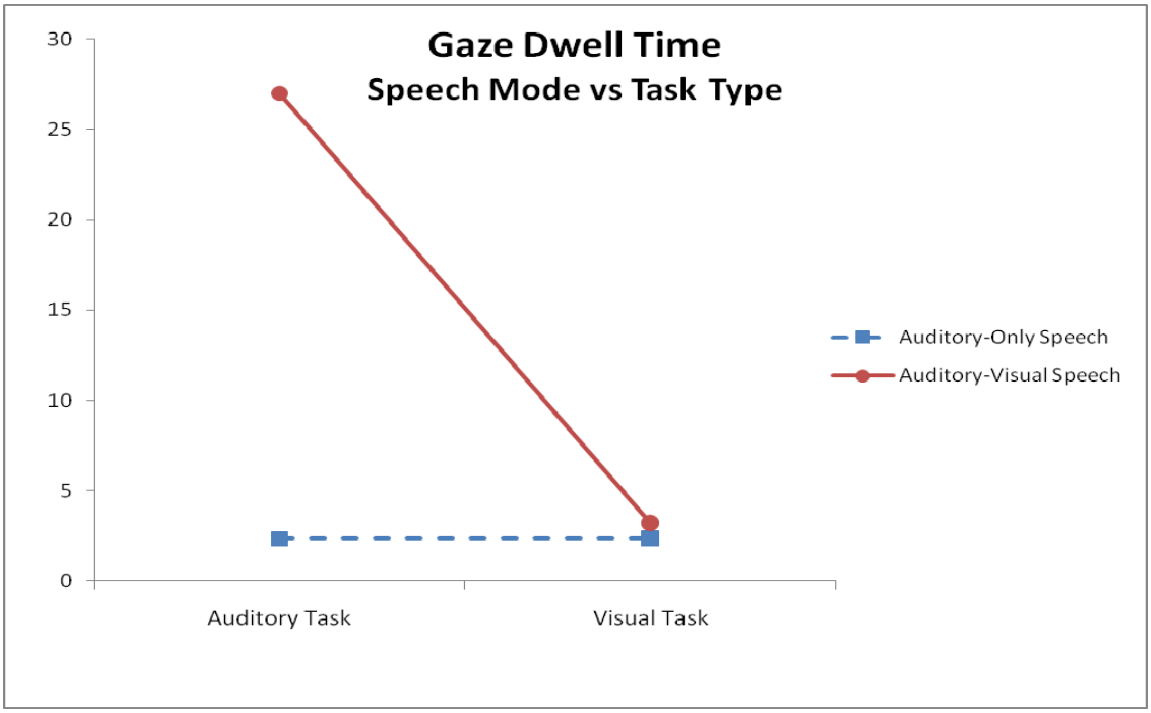

Figure 25. Relationship between Speech Modality and Task Type (Gaze Dwell Times)

V. DISCUSSION

Recalling that the hypothesis being investigated was that the use of a computer-animated facial avatar would improved performance in a multitask scenario that required multimodal processing (visual and auditory), the experimental results indicated that the facial avatar improved performance of verbal tasks under certain conditions. The facial avatar improved speech comprehension during difficult and/or auditory tasks. The facial avatar did not affect the performance of concurrent tasks.

A. WORD IDENTIFICATION

1. Overall

The simple presence of the facial avatar did not have a significant effect on the participants' ability to correctly identify the target word of the verbal tasks. However, significant interactions were identified that involved the presence or absence of the facial avatar. Speech Modality interacted significantly with Task Difficulty, Task Type and Task Difficulty/Type.

2. Speech Modality by Task Difficulty

Recalling Figure 13, the participants' Word Identification scores were lower for the more difficult tasks for the auditory-only presentation of the verbal sentence. For the auditory-visual presentation of the verbal sentence, the Word Identification scores remained fairly constant.

This result indicated that the facial avatar allowed participants to maintain their level of comprehension of speech-in-noise regardless of the difficulty of the concurrent task. This mitigation of the decrement to speech comprehension otherwise associated with increased task difficulty provides support for the incorporation of a facial avatar into communication systems.

3. Speech Modality by Task Type

The participants' Word Identification scores were higher when the facial avatar was present during concurrent auditory tasks and lower when the facial avatar was present during the concurrent visual tasks. This result implied that, although the facial avatar improved comprehension of speech-in-noise during auditory tasks, the presence of the facial avatar interfered with comprehension of speech-in-noise when the visual resources were being otherwise employed.

The interaction between Speech Modality and Task Type is consistent with Wickens' Multiple Resource Theory (2001), keeping in mind that the participants were instructed that the concurrent tasks were their primary tasks.

During the concurrent auditory tasks, the facial avatar provided visual cues that aided in the correct identification of the target word of the verbal messages. With the concurrent auditory task being the primary task it may have required most of the participants' limited auditory resources to listen to the changing tones. The facial avatar allowed the participants to supplement their remaining auditory resources with their visual resources. This verbal processing enhancement allowed the participants to correctly identify the target words more often when the facial avatar was present.

During the concurrent visual tasks, the facial avatar was generally ignored by the participants (as evidenced by the eye tracking data). However, the reduced Word Identification scores suggest that when the facial avatar was present, and could not be directly attended to, the visual speech cues caused confusion in interpreting the verbal message. With the concurrent visual task being the primary task, the remaining visual resources were insufficient to aid in correctly processing the visual speech cues. This interference may have negatively affected the verbal processing, which in turn resulted in lower Word Identification scores during concurrent visual tasks.

4. Speech Modality by Task Type by Task Difficulty

The three-way interaction between Speech Modality, Task Type and Task Difficulty provided the most insight into the effects of the facial avatar on Word Identification during concurrent tasks. Figure 15 provides a visual representation of interaction and displays Word Identification scores by Speech Modality in terms of the type of task, and then by difficulty.

For concurrent auditory tasks, the presentation of the facial avatar improved comprehension of speech-in-noise regardless of the difficulty level. The visual cues provided by the facial avatar increased Word Identification scores by supplementing the limited auditory resources with otherwise unused visual resources.

For the concurrent visual tasks, the relationship between the Speech Modality and the type and difficulty of the concurrent task was more complex. For high difficulty visual tasks, the scores for comprehension of speech-in-noise were nearly identical. During those more difficult concurrent visual tasks the participants' visual resources were engaged to such an extent that little, if any, visual resources were available for the facial avatar. This resulted in similar Word Identification scores regardless of whether the facial avatar was present or not.

However, for less difficult visual tasks the absence of the facial avatar coincided with the increased comprehension. During the less difficult concurrent visual tasks the participants' visual resources were not engaged to the same degree. When the facial avatar was not present there was no interference between the tasks (one purely auditory and one purely visual) and the overall workload was relatively low. This lack of interference and decreased workload resulted in better comprehension of the speech-in-noise.

The three-way interaction between Speech Modality, Task Type and Task Difficulty indicated that a facial avatar may be suitable for improving comprehension of speech-in-noise while concurrent auditory tasks are being

performed, regardless of the difficulty of the auditory task. However, a facial avatar may not be suitable for use during concurrent visual tasks.

B. TASK PERFORMANCE

One of the concerns regarding the presentation of a facial avatar was that it may act as a distraction and reduce the performance of other tasks. Fortunately, the presence or absence of the facial avatar had no significant effect on performance of the concurrent auditory or visual tasks (tone change detection and target icon count, respectively).

Referring back to Table 5, there was no main effect for Speech Modality. Additionally, none of the significant interactions involved Speech Modality. This outcome indicated that the presence of the facial avatar neither improved nor degraded performance of the concurrent tasks. This finding is very important, it alleviated the concern that presenting a facial avatar in an attempt to improve comprehension of speech would interfere with the performance of other tasks.

C. GAZE DWELL TIME

The lack of normality of the data may be attributed to the variety of behaviors exhibited by participants during the auditory tasks. Some participants fixated their gaze on the center of the screen regardless of the presence or absence of the facial avatar. Others closed their eyes or averted their gaze away from the screen. The remainder fixated their gaze on the facial avatar while it was “speaking”. During visual tasks, the participants’ gaze rarely lingered on the facial avatar for more than a brief moment.

The eye tracking data indicated that the participants gazed at the facial avatar primarily during the concurrent auditory tasks. Very little time was spent focused on the facial avatar during the concurrent visual tasks. Participants indicated that visually searching the screen for the target icons prevented them

from making use of the facial avatar. A lower degree of difficulty may have allowed them to divide their time between the visual search task and looking at the facial avatar.

The participants attended to the facial avatar more often when the Sentence Predictability was high. When the verbal message provided few contextual clues to the identity of the target word, there was little utility in attending to the facial avatar during the early portions of the verbal messages. The low predictability sentences had very similar structures, which may have allowed the participants to anticipate their lack of contextual clues within the first few words of the sentence.

Participants were directed to treat the concurrent auditory or visual task as the primary task. If the participants had been permitted to prioritize the tasks as they saw fit, more attention may have been directed towards the facial avatar during the visual tasks. However, it was necessary to control the precedence of the participants' tasks in order to reduce the variability that would have resulted from allowing them to choose. The participants appeared to follow the instruction regarding the priority of tasks, they were not observed actively directing their gaze to the facial avatar during concurrent visual tasks. As well, both the visual and auditory concurrent tasks were not significantly affected by the presence of the facial avatar.

D. REVIEW

When Sumby and Pollack (1954) investigated the usefulness of being able to see a speaker's mouth in a noisy environment, they speculated that augmenting auditory communication with visual cues would prove useful during noisy military operations. The results of this experimental study support their conjecture.

The computer-animated facial avatar used in this study improved speech comprehension under noisy conditions (depending on task type and/or difficulty)

in a manner similar to the animated face employed by Massaro and Cohen (1995). As with the study by Ouni et al. (2007), limiting the avatar to primarily the lips and teeth did not negate its effectiveness.

The performance on the verbal tasks (Word Identification) was consistent with Wickens' Multiple Resource Theory (2001). The cognitive workload associated with the verbal tasks was divided between the auditory and visual resources. The facial avatar improved the comprehension of speech-in-noise during concurrent auditory tasks when some of the workload associated with the verbal task was processed visually. The facial avatar decreased the comprehension of speech-in-noise during the completion of concurrent visual tasks; the visual task interfered with the visual processing of the verbal message.

VI. CONCLUSIONS

The hypothesis investigated was: The use of a computer-animated facial avatar will improve performance in a multitask scenario that requires multimodal processing (visual and auditory).

The primary goal was to determine whether the presentation of a computer-animated facial avatar increased comprehensibility of speech-in-noise while participants performed concurrent tasks. The secondary goal was to determine whether the presentation of a computer-animated facial avatar altered performance on the concurrent tasks.

A. EFFICACY OF THE FACIAL AVATAR

1. Comprehension of Speech-in-Noise

Based simply on the effect of the presence or absence of the facial avatar, the comprehension of speech-in-noise was not significantly improved by the use of the computer-animated facial avatar. However, the presence of the facial avatar did affect the comprehension of speech-in-noise under certain conditions.

There was a significant interaction between the presence of the facial avatar and the difficulty of the concurrent task. The facial avatar was associated with an improvement of the comprehension of verbal messages when the concurrent tasks were at the higher difficulty level.

There was a significant interaction between the presence of the facial avatar and the type of concurrent task. The facial avatar was associated with an improvement of the comprehension of verbal messages during concurrent auditory tasks, and a decrease in the comprehension of verbal messages during concurrent visual tasks.

There was a significant interaction between the presence of the facial avatar, the type of concurrent task and the difficulty of the concurrent task. The

facial avatar was associated with improved comprehension of verbal messages during concurrent auditory tasks, regardless of difficulty level. The facial avatar was associated with decreased comprehension of verbal messages during lower difficulty concurrent visual tasks, but not higher difficulty concurrent visual tasks.

2. Performance of Concurrent Tasks

The presence of the computer-animated facial avatar did not significantly affect the performance of the concurrent auditory or visual tasks.

3. Overall Research Question

The hypothesis that the use of a computer-animated facial avatar will improve performance in a multitask scenario that requires multimodal processing is partially supported. The performance of verbal (listening) tasks is improved under certain conditions; the performance of the concurrent auditory and visual tasks is not affected.

B. RELEVANT DOMAINS OF HSI

1. Human Factors Engineering

The use of a computer-animated facial avatar should prove to be beneficial for improving verbal comprehension in noisy environments, particularly when verbal communication or other auditory tasks are the primary concern of the individual. The facial avatar should act to partially offset the effects of environmental sounds, negating the need for the individual to increase the loudness of the speakers or headset being used. Improved comprehension at lower sound pressure levels will have the added benefit of preventing the individual from contributing unnecessarily to the environmental noise, and may help prevent the ambient noise from reaching levels that could contribute to hearing damage.

Increased comprehension of speech-in-noise will reduce the need for verbal messages to be repeated. The reduction of repetition will reduce overall message traffic, improve the efficiency of the communication system and reduce time lost due to repetition. Reduced message traffic has tactical benefits as well.

More reliable comprehension of the verbal messages should reduce the cognitive workload of both the originator and recipient of the verbal messages. The reduction in workload may have the secondary benefit of reducing stress and mental fatigue.

2. Safety

Improved comprehension of verbal messages should lead to fewer errors and fewer subsequent accidents. Acting upon incorrectly interpreted information may result in incorrect actions being taken, faulty decision making or inaction when action was warranted.

Although the facial avatar did not interfere with the performance of other tasks during this study, judicious use of the facial avatar should prevent it from acting as a distraction and becoming a source of errors and accidents.

3. Training

Instruction is one of the key components of training. Improving the effectiveness of instructions given over a communication system will improve the comprehension of those instructions, and increase student learning. Impaired verbal communication may lead to impaired or incorrect learning.

The use of a facial avatar has the potential to enhance the learning of foreign languages. Foreign languages often possess phonemes unique to that particular language; consequently, learners frequently substitute similar sounding phonemes from their native languages. Although these phonemes sound correct to the learner, they are incorrect nonetheless. The ability to visualize the associated visemes provides the learner with the opportunity to imitate the

correct movements of the lips and tongue. Correct imitation of the physical components of speech improves the likelihood of correctly imitating the phoneme. The acquisition of the foreign language is subsequently faster and more accurate.

4. Personnel

The use of a computer-animated facial avatar has the potential to increase the retention and productive employment of personnel. The incorporation of a facial avatar into a communication system has the potential to include individuals that may have otherwise been excluded from specific roles or tasks. Since the ability to visualize a mouth during speech can serve to offset as much as a 4 to 6 dB of hearing loss (Summerfield, 1992), individuals who possess hearing impairments that marginally prohibit them from being employed in certain roles can potentially be retained in those roles.

C. RECOMMENDATIONS

1. Lessons Learned

In retrospect, the study could have been improved in several respects. Manipulation of the difficulty of the tasks would have reduced the floor and ceiling effects, and would have increased the differences between the Task Difficulty levels.

Because there was no significant interaction between Speech Modality and Sentence Predictability, noise levels could be manipulated instead of Sentence Predictability. This has the potential to provide additional insight into the utility of the facial avatar; it may prove more beneficial as the noise level increases.

Selecting a different visual task, particularly if it has a lower cognitive workload, may have allowed participants to attend to the facial avatar more. If

the participants had the opportunity to actively attend to the facial avatar, very different results for Word Identification and Task Performance may have been observed.

2. Future Research

Based upon the results of this study, future research should focus on employing the facial avatar during concurrent auditory tasks. Additionally, using background noise other than white noise should provide a better indication of the suitability of incorporating computer-animated facial avatars into communication systems. This should provide insight into the potential “real world” applications of the avatar.

Future studies should also examine the minimum level of realism required for the avatar to still be effective. The avatar used during this study utilized a realistic looking mouth at a high frame rate, future studies should investigate the minimum degree of complexity required to improve the comprehension of speech-in-noise. Simpler avatars should require less computer processing power to animate. The avatar can be made simpler by manipulating the frame rate of the animation or the realism of the model (e.g., realistic mouth, “cartoon” mouth or simple line drawing).

3. Potential Application

Because the facial avatar provided the most benefit during concurrent auditory tasks, employing the avatar in roles that involve minimal visual cognitive loads should be the most beneficial. Individuals working in a command center must often monitor multiple radio networks and actively listen to one conversation while several other voices are speaking.

A visual communication display can be created that presents a computer-animated facial avatar for each network being monitored. This would allow the radio operator to focus his visual attention on the avatar associated with the

network in which he is interested. The visual cues provided by the avatar should help the listener selectively attend to the conversation he is primarily concerned with at the moment.

LIST OF REFERENCES

Abel, S. M., & Paik, J. E. S. (2005). Source identification with ANR earmuffs. *Noise & Health. 7*(27), 1–10.

Abel, S. M., Tsang, S., & Boyne, S. (2007). Sound localization with communication headsets: Comparison of passive and active systems. *Noise & Health. 9*(37), 101–107.

Acton, G. S., & Schroeder, D. H. (2001). Sensory discrimination as related to general intelligence. *Intelligence. 29*, 263–271.

Buck, K. (2000). Performance of hearing protectors in impulse noise. *North Atlantic Treaty Organization Research and Technology Organization. EN-11,* 3-1–3-10.

Calvert, A. C., Bullmore, E. T., Beammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*. *276*, 593–596.

Calvert, G. A. & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrate of visible speech. *Journal of Cognitive Neuroscience*. *15*(1), 57–70.

Chen, Y. C., & Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory-visual speech perception. *Journal of the Acoustical Society of America*. *126*(2), 858–865.

Department of Defense. (2010). *Joint communication systems* (Joint Publication 6-0). Washington, DC: U.S. Government Printing Office.

Dyer, J. L., & Tucker, J. J. (2009). *Training analyses supporting the Land Warrior and Ground Soldier Systems* (ARI Research Report 1904). Alexandria, VA: U.S. Army Research Institute for Behavioral and Social Science.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors 37*(1), 32–64.

Giguere, C., Laroche, C., & Vaillancourt, V. (2008). Modelling the effect of personal hearing protection and communication devices on speech perception in noise. *Contract Report CR 2008-178 DRDC Toronto.*

Girin, L., Schwartz, J., & Feng, G. (2001). Audio-visual enhancement of speech in noise. *Journal of the Acoustical Society of America*. *109*(6), 3007–3020.

Kalikow, D. N., Stevens, K. N., & Elliot, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*. *61*(5), 1337–1351.

Licht, D. M., Polzella, D. J., & Boff, K. (1989). *Human factors, ergonomics, and human factors engineering: An analysis of definitions*. CSERIAC-89-01. Wright Patterson AFB, Dayton, OH: CSERIAC.

Lucey, P., Martin, T., & Sridharan, S. (2004). Confusability of phonemes grouped according to their viseme classes in noisy environments. *Proceedings of the 10th Australian International Conference on Speech & Technology.* 265–270.

Massaro, D. W. & Cohen, M. M. (1995). Perceiving talking faces. *Current Directions in Psychological Science*. *4*(4), 104–109.

Nefian, A.V., Liang, L., Pi, X., Liu, X., & Murphy, K. (2002). Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal of Applied Signal Processing. 11*, 1274–1288.

Nicholls, M. E. R., Searle, D. A., & Bradshaw, J. L. (2004). Read my lips: Asymmetries in the visual expression and perception of speech revealed through the McGurk effect. *American Psychological Society*. *15*(2), 138–141.

Ouni, S., Cohen, M. C., Ishak, H., & Massaro, D. W. (2007). Visual contribution to speech perception: Measuring the intelligibility of animated talking heads. *EURASIP Journal of Audio, Speech, and Music Processing. 2007,* 1–12.

Robert-Ribes, J., Schwartz, J., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America*. *103*(6), 3677–3689.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancements of speech in noisy environments. *Cerebral Cortex*. *17*, 1147–1153.

Scharine, A. A., Henry, P. P., & Binseel, M. S. (2005). *An evaluation of selected communication assemblies and hearing protection systems: A field study conducted for the Future Force Warrior Integrated Headgear Integrated Process Team* (ARL-TR-3475). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*. *26*(2), 21–215.

Summerfield, Q. (1992). Lipreading and audio-visual perception. *Philosophical Transactions: Biological Sciences*. *335*(1273), 71–78.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomic Science*. *3*(2), 159–177.

Zieniewicz, M. J., Johnson, D. C., Wong, D. C., & Flatt, J. D. (2002). The evolution of Army wearable computers. *PERVASIVE computing. 4*, 31–40.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California